МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»
Кафедра МОВКиИС

АНАЛИЗ, ОПТИМИЗАЦИЯ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ «БЫСТРЫХ» АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ НА ГРАФОВЫХ СТРУКТУРАХ

АВТОРЕФЕРАТ

НАУЧНО-КВАЛИФИКАЦИОННОЙ РАБОТЫ
(ДИССЕРТАЦИИ)

аспиранта 4 курса
направления 09.06.01 – Информатика и вычислительная техника
направленности «Математическое моделирование, численные методы и комплексы программ»

ИОНКИНА МИХАИЛА СЕРГЕЕВИЧА

Научный руководитель,
д.ф.-м.н., профессор Андрейченко Д. К.

Зав. кафедрой,
д.ф.-м.н., профессор Андрейченко Д. К.

Саратов 2021

ВВЕДЕНИЕ

Актуальность темы

В настоящий момент наблюдается значительный рост объема данных. Ежедневно пользователи Интернета посещают множество сайтов, отмечают на географических картах свое местоположение, обмениваются письмами, мультимедиа файлами, документами, общаются в социальных сетях и т. д. Возникают новые задачи, связанные с обработкой и анализом больших объемов данных, а, следовательно, появляются новые инструменты и технологии для решения таких задач.

Одной из основных задач анализа данных является задача кластеризации (или кластерный анализ) - выделение сообществ (кластеров) разных объектов: пользователей, сайтов, продуктов интернет-магазинов и так далее. Например, с помощью нее можно выявлять схожие производственные ситуации, которые в дальнейшем приводят к браку продукции. Также данный подход позволяет разбивать выборку респондентов на группы в ходе социологических исследований [1]. Методы кластеризации можно также использовать для сегментации изображений (что необходимо для так называемой технологии компьютерного зрения), для распознавания образов и рукописного текста, для извлечения информации и для многого другого [1, 2, 3].

Анализ доступных источников информации показал, что, несмотря на актуальность задачи разбиения набора данных на кластеры, она до сих пор не решена окончательно. Существует огромное множество алгоритмов для решения данной проблемы, только некоторые из них представлены в источниках [1 - 8]. И среди всего многообразия методов нет универсального – у каждого алгоритма есть свои ограничения, преимущества и недостатки. Например, в книгах [3, 6] и на интернет ресурсах [8, 9] утверждается, что некоторым алгоритмам необходимы априорные знания о предполагаемом результате их работы (например, о количестве получаемых кластеров). В

статье [10] авторы приводят варианты решения проблемы плохой масштабируемости этих алгоритмов на большие объемы данных. В публикации [11] авторы разработали иерархический алгоритм, имеющий логарифмическую сложность, тем самым представив способ решения проблемы нелинейного времени выполнения. В статье [12] описывается алгоритм, который позволяет разделять граф на сообщества опираясь не только на связи между вершинами, но и учитывая характеристики этих вершин. В источниках [1, 13] даются рекомендации по применению методов кластеризации к различным типам входных данных (например, ориентированные графы, изображения, тексты).

Также в статье [14] предлагается способ нахождения сообществ с помощью графовых нейронных сетей (Graph Neural Networks). В статье [15] авторы предлагают использовать глубокое обучение для поиска сообществ и показывают, что такой подход превосходит многие из алгоритмов, которые не используют нейронные сети. Также в работе [16] авторы рассматривают глубокое обучение вместе с методом к-средних для поиска сообществ в графах и показывают, что их подход превосходит спектральные методы с точки зрения нормализованной взаимной информации, но при этом проигрывает им во времени выполнения, они предлагают решить эту проблему с помощью распараллеливания выполнения алгоритма к-средних используя Apache Hadoop или Apache Spark. Много работ появилось по нахождению сообществ в динамических сетях [17]. В работе [18] проводится сравнение современных методов поиска сообществ и делается вывод, что вместо того, чтобы искать наиболее подходящий алгоритм для анализа сложных сетей, следует больше внимания уделить пониманию того, о каком свойстве (аспекте, структуре) сети мы хотим получить новые знания. Так, было выяснено, что cut-based методы обеспечивают хорошее разделение на сбалансированные группы, методы кластеризации выполняют разделение на сообщества с высокой внутренней плотностью, стохастические блочные

модели (большой обзор таких моделей представлен в статье [19]) позволяют находить группы, внутри которых находятся наиболее похожие вершины по их свойствам. В работе [20] представлен новый способ нахождения сообществ в социальных сетях, основываясь на информации о пользователях и сообществах других социальных сетей. Похожая задача решается в работе [21] только уже с применением нейронных сетей. В работе [22] предлагается фреймворк, который находит сообщества основываясь на идее о том, как именно формируются группы в социальных сетях. Авторы следующих трех статьей [23, 24, 25] предложили методы поиска сообществ, которые одновременно учитывают и топологическую структуру сети и атрибуты (свойства) ее узлов. В работе [26] представлен алгоритм, который использует идею «центральности» в сети с новой стороны: центральность позволяет отличать лидеров (вершин, соединяющихся с различными сообществами) от последователей (вершин, имеющих соседей только внутри сообщества). Также существуют методы, которые рассматривают сообщества как динамические строительные блоки (dynamical building block) и позволяют находить группы вершин, которые влияют или которые находятся под влиянием какого-либо динамического процесса. Также сети могут иметь совершенно разную структуру: directed [27], temporal [28, 29], multi-layer, multiplex [30]. Видно, что различных алгоритмов огромное количество: от обычных алгоритмов до методов, основанных на построении модели графа и нейронных сетях. Существуют списки методов, с некоторыми из них можно ознакомиться, например, здесь [31, 32].

В ходе анализа доступной литературы и публикаций было обнаружено, что на данный момент не существует наилучшего или универсального формального способа оценки качества разбиения набора данных на кластеры. Существует целый ряд различных эвристических критериев, и каждый из них подходит под определенный тип задачи и данных [1, 27]. Следовательно, для определения качества работы алгоритмов кластеризации на конкретных

данных требуется эксперт в предметной области, который бы мог оценить осмысленность найденных кластеров. Также на результат разбиения на кластеры существенно влияет метрика, то есть мера близости между двумя объектами. Подбор метрики, как правило, субъективен, так как исследователь определяет ее в зависимости от структуры и типа данных, от количества атрибутов и т.д. [1, 33].

В связи с этим возникла потребность в анализе существующих алгоритмов кластеризации на графовых структурах, в выявлении их преимуществ и недостатков.

Целью данной работы является развитие алгоритмов кластеризации на графовых структурах, а также выполнение сравнительного анализа алгоритмов поиска сообществ в графах.

Для достижения поставленной цели решались следующие задачи:

1. Описать основные понятия и определения больших данных, data mining и машинного обучения и, в частности, задачи кластеризации;
2. Изучить наиболее популярные алгоритмы кластеризации (поиска сообществ) в графах, кратко описать идеи и механизмы их работы. Найти существующие реализации этих алгоритмов или реализовать самостоятельно;
3. Описать основные способы поиска расстояний между объектами применимые для задачи кластеризации;
4. Рассмотреть наиболее популярные методы оценки качества кластеризации;
5. Для графов со взвешенными вершинами улучшить алгоритм «Louvain» применяемый для поиска сообществ;
6. Подобрать наглядные примеры и данные с уже известными (ground-truth) сообществами для сравнительного анализа выбранных алгоритмов;
7. Выполнить анализ результатов работы алгоритмов «вручную», то есть, не используя каких-либо метрик и автоматических средств проверки

качества;

8. Выполнить анализ результатов работы алгоритмов, используя наиболее популярные метрики и функционалы качества;

9. Выявить общие признаки наборов данных, представляющих собой графовые структуры;

10. Исследовать алгоритмы поиска сообществ в графах на предмет принципов их работы;

11. Классифицировать графовые наборы данных и алгоритмы поиска сообществ;

12. Выполнить сопоставительный анализ классов графов и алгоритмов с точки зрения качества результатов разбиения на кластеры в сравнении с ground-truth сообществами.

Научная новизна

В ходе данного исследования в алгоритме Louvain для кластеризации графов была обнаружена новая проблема (противоположная известной проблеме resolution limit для больших графов), а именно, – разбиение малого набора данных на слишком мелкие (относительно всего графа) сообщества. В данной работе предложен вариант решения путем изменения формулы оптимизации модулярности в алгоритме Louvain.

В работе выполнена эффективная реализация алгоритма Smart Local Moving на языке программирования Python, что позволяет использовать этот алгоритм в широком круге проектов, посвященных анализу данных.

Реализована библиотека на языке программирования Python для выгрузки и построения графа связей пользователей социальной сети «ВКонтакте».

Реализована библиотека на языке программирования Python для выполнения сравнительного анализа работы алгоритмов поиска сообществ в графах.

Также представлена классификация графовых наборов данных, представляющих собой модели объектов реального мира, на три крупные группы по типу построения графа.

Представлена классификация методов поиска сообществ по принципам (свойствам), лежащим в основе работы этих методов.

Приведен сопоставительный анализ классов графов и классов алгоритмов с точки зрения качества результатов разбиения на кластеры в сравнении с ground-truth сообществами.

Практическая значимость

Полученные результаты могут быть использованы организациями и специалистами в области интеллектуального анализа данных для выявления сообществ в наборах данных, представляющих собой графовые структуры. В последнее время для решения подобных задач используется язык программирования Python. В данной работе представлена реализация эффективного алгоритма Smart Local Moving на этом языке, что позволяет использовать этот алгоритм в большем количестве проектов. Также реализована библиотека для выгрузки и построения графа связей пользователей социальной сети «ВКонтакте». А реализованная библиотека для сравнительного анализа алгоритмов позволяет производить тесты на любых других наборах данных, представляющих собой графовые структуры.

Представленная классификация графов и алгоритмов и их сопоставительный анализ, позволяют значительно сократить время и ресурсы, затрачиваемые специалистами на поиск подходящего алгоритма для конкретного набора данных.

Также материалы работы могут быть использованы в высшей школе в процессе обучения бакалавров, специалистов и магистров в лекционных курсах по машинному обучению и анализу данных.

Структура и объём работы. Научно-квалификационная работа (диссертация) состоит из введения, 4 разделов, заключения, списка

использованных источников и 9 приложений. Общий объем работы – 122 страницы, из них 91 страница – основное содержание. В работе 34 рисунка и 9 таблиц, а список использованных источников информации содержит 62 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Интеллектуальный анализ данных и машинное обучение» посвящен обзору основных понятий, связанных с интеллектуальным анализом данных и машинном обучением. Рассмотрено понятие «большие данные» и некоторые техники анализа, применимые к ним. Приведена краткая история возникновения термина «data mining», дана общая постановка задачи data mining и классификация таких задач. Приведено формальное определение задачи обучения по прецедентам (обучение с учителем) и задачи кластеризации (обучение без учителя).

Второй раздел «Поиск сообществ в неориентированных графах» посвящен описанию одной из задач машинного обучения, а именно, - поиску сообществ в неориентированных графах. В начале раздела описывается понятие «сообщество в графе» и рассматриваются проблемы, которые возникают из-за того, что не существует формального определения понятия «сообщество в графе». Затрагивается проблема оценки результатов работы алгоритмов. Далее рассматриваются самые популярные на данный момент алгоритмы для поиска сообществ: Infomap, Walktrap, Label Propagation, Fastgreedy, Edge Betweeness, Louvain, Smart Local Moving. Делается вывод о том, что неопределенность постановки задачи поиска сообществ в графах приводит к разнообразию подходов к решению поставленной задачи и к разнообразию методов оценивания алгоритмов.

Третий раздел «Метрики и оценка качества кластеризации» посвящен более подробному рассмотрению проблемы оценки качества кластеризации. Здесь описываются некоторые стандартные метрики, применяемые в алгоритмах кластеризации, дается определение понятий

«функция потерь» и «функционал качества», описывается «модулярность» - наиболее популярный функционал качества для задачи поиска сообществ в графах, а также рассматривается «нормализованная взаимная информация» (NMI) - популярный в последнее время способ сравнения разных разбиений. Приводятся преимущества и недостатки модулярности и нормализованной взаимной информации.

Четвертый раздел «Практическая часть» посвящен практическому применению рассмотренных алгоритмов, оценке их характеристик и проведению сравнительного анализа.

Один из самых известных на данный момент алгоритмов поиска сообществ Louvain использует в своей основе модулярность. Обычно алгоритмы, использующие модулярность не замечают мелкие (относительно всего графа) сообщества и объединяют их в одно сообщество. Эта проблема получила название resolution limit. Существует несколько работ, посвященных исследованию этой проблемы и способам ее решения (например, в [34, 35, 36, 37]). В ходе экспериментов выяснилось, что помимо обозначенной выше проблемы, на небольших объемах данных алгоритм Louvain делит их на слишком мелкие сообщества. И можно сказать, что эта проблема является прямо противоположной проблеме resolution limit. Для решения этой проблемы было предложено учитывать веса вершин графа при подсчете модулярности путем изменения формулы оптимизации модулярности в алгоритме Louvain. Был проведен эксперимент, в результате которого, расставив веса вершин у графа, удалось сократить количество кластеров в получающемся разбиении и сделав таким образом разбиение более качественным.

Далее было проведено сравнение алгоритмов Infomap, Walktrap, Label Propagation, Fastgreedy, Edge Betweeness, Louvain на «искусственных» данных. Первый набор данных представлял собой граф, состоящий из двух не связанных между собой циклов, а второй набор – известный граф «клуб

карате». Сравнение производилось с помощью модулярности и «вручную» (то есть оба набора данных были заранее разбиты на сообщества, которые в дальнейшем сравнивались с разбиениями, получающимися в результате работы алгоритмов). В результате выполнения двух тестов видно, что результаты работы алгоритмов зависят от структуры графа. Например, алгоритм Fastgreedy показал лучший результат при анализе графа «клуб карате», но при этом плохо справился с анализом графа, состоящего из двух циклов, не соединенных между собой. Но, например, результаты таких алгоритмов как Infomap и Louvain были приемлемого качества в этих двух тестах. Также видно, что решение, лучшее с нашей точки зрения, не всегда совпадает решением, лучшим с точки зрения функционала качества. Отсюда следует, что необходимо учитывать принцип работы функционала. Например, сети с высокой модулярностью имеют плотные связи между вершинами внутри сообществ, но редкие связи между вершинами из разных сообществ. Но ведь не всегда плотные связи внутри одного сообщества означают, что его вершины действительно должны ему принадлежать.

В ходе анализа доступной литературы и публикаций было обнаружено, что алгоритм Smart Local Moving (SLM) обладает преимуществом по сравнению с известным алгоритмом Louvain: он позволяет находить разбиение графа с большей модулярностью, чем Louvain. Поэтому было решено реализовать данный алгоритм на языке программирования Python, чтобы была возможность выполнять сравнительные тесты с другими уже реализованными алгоритмами из библиотеки igraph в одной среде. Это позволит не учитывать в ходе анализа, например, различные механизмы сборки мусора языков программирования, время, затрачиваемое на компиляцию или интерпретацию исходного кода, внутреннюю реализацию базовых структур данных языков и т. д. В ходе реализации использовалось описание алгоритма SLM и его псевдокод (приложение А) из статьи [38].

После анализа работы всех перечисленных выше алгоритмов на «искусственных» примерах следующим очевидным шагом является выполнение этого же анализа, но уже на «реальных» данных. Также было решено сравнить получившиеся сообщества каждого отдельного алгоритма с ground-truth сообществами, то есть с уже известными сообществами этой сети, сформированными по какому-либо признаку (например, группа в социальной сети, в которой состоят поклонники какого-нибудь артиста).

В качестве таких данных были взяты общедоступные данные трех сайтов (You Tube, Amazon и Live Journal) [39], которые представляют собой большие разреженные графы. Все данные «обезличены» и представляют собой просто набор идентификаторов. Было проведено сравнение результатов работы алгоритмов по их времени выполнения и по значению модулярности отдельно для каждого набора данных. Время выполнения всех алгоритмов совпало с теоретическими оценками их времени выполнения. С помощью метрики «нормализованная взаимная информация» было проведено сравнение получившихся разбиений с ground-truth сообществами. Так как у некоторых алгоритмов мы получили примерно одинаковые значения модулярности и NMI, то необходимо было выяснить получаются ли разбиения одинаковыми или это другие разбиения, но с похожей модулярностью. Для этого было проведено сравнение разбиений алгоритмов между собой с помощью метрики NMI. Это показало, например, что алгоритмы Louvain и SLM, а также Walktrap и Label Propagation находят наиболее похожие разбиения, не зависимо от набора данных.

Так как предыдущий анализ выполнялся на «обезличенных» данных, то не совсем понятно какой смысл несут в себе получившиеся разбиения. Поэтому для дальнейшего анализа был решено использовать данные социальной сети «ВКонтакте» (далее по тексту – набор данных Vk). Для этого была реализована библиотека на языке Python для загрузки и

предварительной обработки информации о пользователях этой социальной сети.

Рассматриваемый граф является моделью отношения “является другом” среди пользователей социальной сети. Был взят один пользователь социальной сети и список всех его друзей, и построен неориентированный граф следующим образом: вершинами являются друзья этого пользователя, и между двумя вершинами проводится ребро, если эти два человека являются друг другу друзьями. Далее получившийся граф был вручную поделен на ground-truth сообщества. Был проведен анализ результатов работы алгоритмов на этом наборе данных с помощью модулярности и NMI. Было проведено сравнение получившихся разбиений с ground-truth сообществами и между собой. В данном примере алгоритмы Louvain и SLM смогли достаточно точно определить социальные группы среди друзей пользователя, а наиболее похожие получились разбиения у алгоритмов Walktrap, Fastgreedy, Edge Betweenness и Infomap и их разбиения довольно сильно отличаются от ground-truth сообществ.

Был сделан вывод, что алгоритмы способны, основываясь только на знании о структуре графа, разбивать его на сообщества, которые несут в себе ценную информацию, которая не была доступна ни исследователю, ни самим алгоритмам. А в этом и заключается основная идея data mining.

На данный момент не существует формального определения понятия «сообщество» в графе. Сейчас сообществом называют набор вершин, связанных между собой каким-либо общим свойством [40]. А значит один и тот же граф может иметь несколько «правильных» разбиений, всё зависит от того, на какой вопрос требуется получить ответ. И в ходе практических экспериментов было обнаружено, что тип графа влияет на последующую интерпретацию результатов работы алгоритмов и практическую применимость результатов разбиения. В связи с этим была проведена классификация графовых наборов данных, представляющих собой модели

объектов реального мира, на три крупные группы по типу построения графа: граф связей, графы, представляющие собой контент, создаваемый самими пользователями, и графы взаимодействия. Также представлена классификация методов поиска сообществ по принципам (свойствам), лежащим в основе работы этих методов: похожесть вершин, плотность ребер, расстояние между вершинами. Приведен сопоставительный анализ классов графов и классов алгоритмов с точки зрения качества результатов разбиения на кластеры в сравнении с ground-truth сообществами. В ходе исследования была обнаружена взаимосвязь между типом графа и типом алгоритма, который применяется для нахождения сообществ в этом графе.

ЗАКЛЮЧЕНИЕ

В работе проведен анализ наиболее популярных алгоритмов поиска сообществ в графах, изучены их особенности выполнения и свойства, дано описание их работы, подобраны необходимые библиотеки для языка программирования Python.

Было приведено сравнение результатов кластеризации алгоритма Louvain с «ручной» обработкой данных. Также для алгоритма Louvain предложен вариант решения проблемы, противоположной проблеме resolution limit, а именно, – проблемы разбиения графа на слишком мелкие (относительно всего графа) сообщества. Для этого в формуле оптимизации модулярности было предложено учитывать веса вершин графа, что помогло подсчитывать модулярность более точно и, таким образом, производить более качественное разделение графа на сообщества.

Реализована библиотека на языке программирования Python для выполнения сравнительного анализа работы алгоритмов поиска сообществ в графах.

Проанализирована работа выбранных алгоритмов на двух тестах (первый тест – на графе, состоящем из двух циклов, второй тест – на графе, который является один из самых известных наборов данных «клуб карате»).

Анализ качества работы алгоритмов осуществлялся «вручную», то есть без использования функционалов качества.

Была представлена эффективная реализация на языке программирования Python алгоритма Smart Local Moving, который является улучшением алгоритма Louvain в максимизации модулярности. Данная реализация позволяет использовать этот алгоритм в широком круге проектов, посвященных анализу данных.

Был проведен сравнительный анализ выбранных алгоритмов на трех наборах данных с ground-truth сообществами, содержащими несколько сотен тысяч вершин и ребер, представляющими собой большие разреженные графы. Оценивались такие характеристики как время выполнения алгоритмов, показатели модулярности и нормализованной взаимной информации (NMI).

Так как данные этих трех наборов были «обезличены», то было решено провести подобный анализ на данных социальной сети «ВКонтакте». Была реализована библиотека на языке программирования Python для выгрузки и построения графа связей пользователей социальной сети «ВКонтакте». Все конкретные выводы и результаты относительно работы алгоритмов представлены в соответствующих главах данной работы. Обобщив, можно сказать, что каждый алгоритм имеет свои преимущества и недостатки. Например, по времени выполнения и максимизации значения модулярности явными лидерами являются алгоритмы Louvain и Smart Local Moving. При этом они не всегда хорошо находят ground-truth сообщества. С этой задачей в двух тестах хорошо справился алгоритм Fastgreedy, но не смог найти «правильную» структуру для данных социальной сети «ВКонтакте».

Также была проведена классификация графовых наборов данных, представляющих собой модели объектов реального мира, на три крупные группы по типу построения графа. А методы поиска сообществ были разделены на несколько категорий, по принципам (свойствам), лежащим в

основе работы этих методов. Был проведен сопоставительный анализ классов графов и классов алгоритмов с точки зрения качества результатов разбиения на кластеры в сравнении с ground-truth сообществами. В результате была обнаружена взаимосвязь между типами графов и типами алгоритмов, что поможет значительно сократить время и ресурсы исследователей прикладных задач по анализу графовых структур.

В ходе исследования был сделан вывод, что, несмотря на актуальность задачи поиска сообществ, она до сих пор не решена окончательно. И среди представленных решений нет универсального – каждый имеет свои ограничения, преимущества и недостатки.

На основе результатов исследования можно реализовать библиотеку, которая позволяла бы в автоматическом режиме подбирать наиболее эффективные алгоритмы поиска сообществ в зависимости от типа графа и типа оптимизационной метрики. Также перспективным направлением для дальнейших исследований в рамках данной работы является анализ алгоритмов, основанных на нейронных сетях.

ОСНОВНЫЕ НАУЧНЫЕ ПУБЛИКАЦИИ ПО ТЕМЕ НАУЧНО-КВАЛИФИКАЦИОННОЙ РАБОТЫ (ДИССЕРТАЦИИ)

1. Ионкин М. С., Огнева М. В. Программная реализация, анализ эффективности и оценка качества алгоритмов кластеризации графовых моделей социальных сетей // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2017. Т. 17, вып. 4. С. 441-451. DOI: 10.18500/1816-9791-2017-17-3-441-451.

2. Ионкин М. С., Огнева М. В. Непрерывная подготовка специалистов в области анализа данных // Электронные образовательные технологии – пространство неограниченных возможностей: материалы I Междунар. науч.-практ. конф. Новосибирск: Изд-во СГУПСа, 2017. – с. 49-54.

3. Ионкин М. С., Огнева М. В. Анализ и классификация алгоритмов кластеризации с точки зрения применимости к различным типам графов // Материалы Международной научной конференции "Компьютерные науки и информационные технологии" памяти А.М.Богомолова. Саратов: Издат. центр «Наука», 2018. – с.167-170.

4. Ionkin M. S., Ogneva M. V. Comparative analysis of community detection algorithms for undirected graphs // Представляем научные достижения миру. Естественные науки : материалы IX научной конференции молодых ученых «Presenting Academic Achievements to the World». Саратов: Изд-во «Саратовский источник», 2019. – Вып. 8. – с. 46-51.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Aggarwal, C. C. Data clustering. Algorithms and applications / Charu C. Aggarwal, Chandan K. Reddy. N.-Y.: Chapman and Hall/CRC, 2014. — 652 p.
2. Jain, A. K., Murty M. N., Flynn, P. J. Data clustering: a review // Acm computing surveys. 1999. vol. 31, № 3. pp. 264-323.
3. Сегаран, Т. Программируем коллективный разум. / Т. Сегаран. – пер. с англ. – Спб: Символ-плюс, 2008. – 368 с.
4. Федоренко, Ю. С. Кластеризация данных на основе нейронного газа и марковских алгоритмов // Молодежный научно-технический вестник. 2014. № 8.
5. Newman, M. E. J. Detecting community structure in networks // The European Physical Journal B - Condensed Matter and Complex Systems. 2004. Volume 38, issue 2. pp. 321–330.
6. Leskovec, J. Mining of massive datasets / Jure Leskovec, Anand Rajaraman, Jeff Ullman; 2nd edition. Cambridge University Press, 2014. – 511 p.
7. Fortunato, S. Community detection in graphs // Physics Reports. 2010. № 486(3). pp. 75-174.

8. Информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс] : [сайт]. URL: www.machinelearning.ru (дата обращения 12.12.2017). Загл. с экрана.
9. Чубукова, И. А. Курс лекций «data mining» // Интернет-университет информационных технологий [Электронный ресурс] : [сайт]. URL: www.intuit.ru/department/database/datamining (дата обращения 02.02.2017). Загл. с экрана.
10. Arnau Prat-Pérez, David Dominguez-Sal, Josep-Lluis Larriba-Pey. High quality, scalable and parallel community detection for large real graphs // Proceedings of the 23rd international conference on World Wide Web. 2014. pp. 225-236.
11. Clauset, A., Newman, M. E. J., Moore, C. Finding community structure in very large networks // Physical Review. 2004. № 70(066111).
12. Jaewon Yang, Julian Mcauley, Jure Leskovec. Community detection in networks with node attributes // IEEE 13th International Conference on Data Mining. 2013. pp. 1151-1156.
13. Witten I. H., Frank E., Hall M. A. Data mining: practical machine learning tools and techniques / Ian H. Witten, Eibe Frank, Mark A. Hall; 3rd edition - San Francisco: Morgan Kaufmann Publishers Inc., - 2011. – pp. 665.
14. Zhengdao Chen. Supervised community detection with line graph neural networks [Электронный ресурс] / Zhengdao Chen, Joan Bruna, Lisha Li // arXiv.org. — Режим доступа : https://arxiv.org/pdf/1705.08415.pdf (дата обращения : 12.04.18).
15. Liang Yang, Learning Xiaochun Cao, Dongxiao He, Chuan Wang, Xiao Wang, Weixiong Zhang. Modularity based Community Detection with Deep Learning // Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. 2016. pp 2252–2258.

16. Vilcek A. Deep learning with K-Means applied to community detection in networks [Электронный ресурс]. URL: http://web.stanford.edu/class/cs224w/projects.html (дата обращения: 10.03.2018).
17. Peizhuo Wang, Lin Gao, Xiaoke Ma. Dynamic community detection based on network structural perturbation and topological similarity // Journal of Statistical Mechanics: Theory and Experiment. 2017. no. 1, p. 013401.
18. Martin Rosvall. Different approaches to community detection [Электронный ресурс] / Martin Rosvall, Jean-Charles Delvenne, Michael T. Schaub, Renaud Lambiotte // arXiv.org. — Режим доступа : https://arxiv.org/pdf/1712.06468.pdf (дата обращения : 12.04.17).
19. Emmanuel Abbe. Community Detection and Stochastic Block Models: Recent Developments // Journal of Machine Learning Research. 2018. Volume 18, no 177. pp. 1-86.
20. Qianyi Zhan, Jiawei Zhang, Philip Yu, Junyuan Xie. Community detection for emerging social networks // World Wide Web. 2017. Volume 20, issue 6. pp 1409–1441.
21. Jie Tang, Tiancheng Lou, Jon Kleinberg, Sen Wu. Transfer Learning to Infer Social Ties across Heterogeneous Networks // ACM Transactions on Information Systems. 2016. Volume 34, issue 2, article no 7. pp 1–43.
22. Wei Chen, Zhenming Liu, Xiaorui Sun, Yajun Wang. Community Detection in Social Networks through Community Formation Games // Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence. 2011. pp. 2576-2581.
23. Liyuan Liu, Linli Xu, Zhen Wangy, Enhong Chen. Community detection based on structure and content: a content propagation perspective // IEEE International Conference on Data Mining. 2015. pp. 271–280.

24. Boujlaleb L., Mammass D., Idarrou A., Idrissa S. Community detection in mobile social networks // Global journal of engineering science and researches. 2015. C. 50-55.
25. Haithum Elhadi, Gady Agam. Structure and attributes community detection benchmark and a novel selection method // Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining. Нью-Йорк, 2013. С. 1474–1476.
26. Devavrat Shah. Leaders, Followers, and Community Detection [Электронный ресурс] / Dhruv Parthasarathy, Devavrat Shah, Tauhid Zaman // arXiv.org. — Режим доступа : https://arxiv.org/pdf/1712.06468.pdf (дата обращения : 12.04.17).
27. Fragkiskos D. Malliaros, Michalis Vazirgiannis. Clustering and community detection in directed networks: a survey // Physics Reports. 2013. Volume 533, issue 4. pp. 95-142.
28. Petter Holme. Temporal Networks [Электронный ресурс] / Petter Holme, Jari Saramäki // arXiv.org. — Режим доступа : https://arxiv.org/pdf/1108.1780.pdf (дата обращения : 12.04.17).
29. Sekara V., Stopczynski A., Lehmann S. Fundamental structures of dynamic social networks // Proceedings of the National Academy of Sciences. 2016. pp. 9977–9982.
30. Boccaletti S., Bianconi G., Criado R., Del Genio, C.I., Gómez-Gardeñes J., Romance M., Sendiña-Nadal I., Wang Z., Zanin, M. The structure and dynamics of multilayer networks // Physics Reports. 2014. Volume 544. pp. 1–122.
31. Lab41. Market Survey: Community Detection // GitHub [Электронный ресурс]: [сайт]. URL: http://lab41.github.io/survey-community-detection (дата обращения 06.09.2018). Загл. с экрана.
32. RapidsAtHKUST. Community Detection Related (Mainly Social Network) // GitHub [Электронный ресурс] : [сайт]. URL:

https://github.com/RapidsAtHKUST/CommunityDetectionCodes/blob/master/Survey/Community-Detection-Survey.md (дата обращения 06.09.2018). Загл. с экрана.

33. Tan, P.-N., Steinbach, M., Vipin, K. Introduction to Data Mining / Pang-Ning Tan, Michael Steinbach, Vipin Kumar, - 1st edition - Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc. – 2005. - pp. 725.
34. Santo Fortunato, Marc Barthélemy. Resolution limit in community detection // Proceedings of the National Academy of Sciences. 2007. № 104. pp 36-41.
35. Kumpula, J. M., Saramaki, J., Kaski, K., Kertesz, J. Limited resolution and multiresolution methods in complex network community detection // Fluctuation Noise Letters. 2007. № 7 (209).
36. Ronhovde, P., Nussinov, Z. Local resolution-limit-free potts model for community detection // Physical Review. 2010. № 81.
37. Traag, V. A., Dooren, P. V., Nesterov, Y. Narrow scope for resolution-limit-free community detection. 2011. Physical Review. № 84.
38. Waltman, L., Van Eck, N.J. A smart local moving algorithm for large-scale modularity-based community detection // The European Physical Journal B. 2013. Volume 86 (11).
39. Jure Leskovec. Stanford Large Network Dataset Collection // Stanford Network Analysis Project [Электронный ресурс] : [сайт]. URL: https://snap.stanford.edu/data (дата обращения 11.02.2018). Загл. с экрана.
40. Ионкин М. С., Огнева М. В. Программная реализация, анализ эффективности и оценка качества алгоритмов кластеризации графовых моделей социальных сетей // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2017. Т. 17, вып. 4. С. 441–451.