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

Кафедра информатики и программирования

РЕАЛИЗАЦИЯ И СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ СОЦИАЛЬНЫХ СЕТЕЙ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Павловой Александры Сергеевны

Научный руководитель,
к.ф.-м.н., доцент _____ М.В. Огнева
подпись, дата

Зав. кафедрой ИиП,
к.ф.-м.н., доцент _____ М.В. Огнева
подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность темы.

В современном мире есть огромное количество информации, которую возможно представить в виде объектов и отношений между ними. Например, объектами могут быть научные публикации. В таком случае, если одна из публикаций ссылается на другую, то между ними есть связь. Таким образом, все имеющиеся научные публикации можно представить в виде графа. Аналогичное представление возможно и для большинства других структур из разнообразных областей знания: люди и их социальные взаимоотношения, ветки метро и станции пересадок, атомы молекулы и химические связи между этими атомами. Для многих подобных структур считается логичным их представление в виде графа. С каждым днем масштабы и сложность таких графов растут и даже такая первостепенная задача, как визуализация, оказывается более чем проблематичной. Одним из возможных подходов по решению этой проблемы является кластеризация. Суть этого подхода заключатся в более высокоуровневом представлении начального графа, т.е. изображении вместо вершин графа группы вершин.

Более того, при помощи кластеризации можно определять группы пользователей со схожими предпочтениями, что, на следующем этапе, помогает выяснить, какая информация будет для этих групп более востребована и интересна. Различные методы кластеризации применяются в задачах машинного обучения с целью понижения размерности графа, в маркетинговых исследованиях для обеспечения оптимизации рекламной и логистической деятельности, в области компьютерного зрения и т.д. Таким образом, задача кластеризации, т.е. выделение сообществ (кластеров) разных объектов, представляет собой одну из основных задач анализа данных.

Несмотря на значимость задачи кластеризации, она на сегодняшний день не решена окончательно. Существуют многочисленные алгоритмы для решения

этой проблемы, но каждый из алгоритмов имеет особые ограничения, достоинства и недостатки.

В связи с этим актуальной является задача оптимизации, анализа и сравнения алгоритмов и их оценок, выявления их преимуществ и недостатков, что, возможно, в дальнейшем приведет к созданию универсального алгоритма.

Цель бакалаврской работы – изучить алгоритмы кластеризации и провести их сравнительный анализ.

При выполнении данной работы были поставлены **следующие задачи**:

1. Изучить теоретический материал, относящийся к алгоритмам кластеризации и методам оценки их работы.
2. Протестировать алгоритмы кластеризации LabelPropogation, Infomap, Walktrap, Louvain на различных обезличенных датасетах с ground truth разбиением.
3. Получить данные для построения графа социальной сети реально существующего пользователя социальной сети «В Контакте» и протестировать алгоритмы кластеризации на полученном графе.
4. Реализовать алгоритм кластеризации Louvain на языке программирования Python и сравнить результат его работы с результатом работы эталонного алгоритма Louvain, представленного в библиотеке igraph.
5. Описать проблему оценивания качества кластеризации, посчитать оценки на всех имеющихся датасетах, сравнить полученные данные.

Структура и объём работы. Бакалаврская работа состоит из введения, четырех разделов, заключения, списка использованных источников и четырех приложений. Общий объем работы – 56 страниц, из них 38 страниц – основное содержание, включая 4 рисунка и 9 таблиц, список использованных источников информации – 23 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Проблема кластеризации и способы ее решения» посвящен последним крупным исследованиям в данной области.

Широкая доступность Интернета стимулирует активность пользователей в социальных сетях. Обнаружение сообществ является одной из важных задач для понимания поведения и функционала таких реальных сетей.

Математически проблема обнаружения сообщества была смоделирована как задача оптимизации, и для ее решения были применены различные метаэвристические подходы. За последнее десятилетие постепенно было разработано множество новых алгоритмов, для решения разнообразных задач оптимизации. Так же стоит отметить, что последние пару лет алгоритмы кластеризации все еще продолжают совершенствовать и применять, как вспомогательный инструмент. Приведем несколько примеров таких исследований.

Биологические, физические и химические процессы, происходящие с млекопитающими в природе, побуждают исследователей предлагать новые алгоритмы кластеризации и использовать их для различных применений. Из 1240 видов летучих мышей большинство ориентируются в пространстве и охотятся посредством эхолокации — испускания звуковых сигналов и регистрации волн, отраженных от окружающих предметов. Именно этот факт и вдохновил Х. Янга в 2020 году разработать новый метаэвристический алгоритм, получивший название Bat Algorithm . Мирджалили и др. представили двоичный алгоритм bat для решения задачи дискретной оптимизации. Успех алгоритма bat вдохновил исследователей на его изучение для решения различных задач оптимизации.

Поскольку масштаб сложных сетей постоянно увеличивается, для повышения эффективности работы происходит усовершенствование уже существующих алгоритмов. Так, например, в ноябре 2021 года был предложен улучшенный быстрый алгоритм Louvain . Алгоритм оптимизирует итеративную

логику от циклической итерации к динамической итерации, что ускоряет скорость сходимости и разбивает локальную древовидную структуру в сети. Разделенная сеть разделяется итеративно, затем древовидная структура добавляется к результатам разбиения, и результаты оптимизируются для сокращения вычислений. Быстрый алгоритм Louvain имеет более высокую агрегацию сообщества, поэтому эффект обнаружения сообщества улучшается. После экспериментального тестирования нескольких групп данных было выяснено, что быстрый алгоритм Louvain превосходит традиционный.

Второй раздел «Обзор алгоритмов кластеризации» включает в себя рассмотрение основных определений и подробное описание основных алгоритмов кластеризации.

Кластеризация (англ. cluster analysis) — задача группировки множества объектов на подмножества (кластеры) таким образом, чтобы объекты из одного кластера были более похожи друг на друга, чем на объекты из других кластеров по какому-либо критерию.

Граф (простой) - пара $G = (V; E)$, где $V \neq \emptyset$ – конечное множество вершин, E – множество пар вершин.

Далее за n будем обозначать количество вершин в графе, а за m — количество ребер.

Если E – множество неупорядоченных пар вершин, то граф называется неориентированным (элементы множества E называются ребрами), иначе – ориентированным (элементы множества E – дуги).

Кластеризацией графа называется разделение множества вершин графа на подмножества, называемые кластерами (сообществами). Характеристикой качества кластеризации является то, что каждый кластер слабо связан с другими кластерами, при этом вершины кластера сильно связаны между собой.

На данный момент существуют около 15 алгоритмов кластеризации, которые отличаются принципом и временем работы, а также эффективностью и устойчивостью. Рассмотрим более подробно следующие: Edge Betweenness,

Fastgreedy, LabelPropogation, Walktrap, Infomap, Louvain и Smart Local Moving. Первые пять — остаточно популярные алгоритмы, реализованные во всевозможных библиотеках для анализа данных. Но алгоритмы Louvain и Smart Local Moving были реализованы относительно недавно (2008 и 2013 гг.). Главным отличием этих алгоритмов от первых пяти является то, что в соответствии с теоретическими оценками они работают намного быстрее.

Рассмотрим каждый алгоритм немного детальнее.

Edge Betweenness — это алгоритм, работающий на основе коэффициента «центральности по посредничеству» (Betweenness), который, в свою очередь, определяется, как количество кратчайших путей между всеми парами вершин, проходящих через данное ребро.

Данный метод имеет значительное количество модификаций, которые в большинстве случаев сводятся к подсчету иных реберных коэффициентов или замене модулярности на другой схожий функционал. Главный недостаток Edge Betweenness — очень большое время работы. Это можно объяснить тем, что подсчет коэффициентов на ребрах является вычислительно сложной задачей. Сложность метода $O(m^2n)$.

Fastgreedy — алгоритм, базирующийся на жадной оптимизации функции модулярности.

Данный алгоритм считается быстрым и вычислительно простым, что способствует применению его для достаточно больших графов. Так же в этот алгоритм достаточно легко добавить заранее известную информацию о составе кластеров, к примеру, если известно, что какие-то определенные вершины должны находиться в одном кластере. Тем не менее, алгоритм включает в себя все недостатки, свойственные жадным методам, и сходится не к самому лучшему решению. Часто метод порождает одно большое сообщество с большинством вершин графа в нем и множество маленьких. Сложность метода $O(mn)$.

LabelPropogation — метод, заключающийся в присвоении меток к каждой вершине. Каждый раз берется метка с самой большой встречаемостью среди смежных вершин.

Метод отличается своей простотой и интуитивностью, но при этом считается вычислительно эффективным. Однако, он склонен выдавать разные результаты и является в каком-то смысле неустойчивым алгоритмом, что заметно на практике. В среднем, этот метод работает похуже остальных. Вычислительная сложность метода LabelPropogation фактически линейная и сравнима с O(m).

Infomap — метод поиска сообществ, основанный на случайном блуждании, информационных потоках в сетях, кодировании и сжатии информации.

Дискретный процесс случайного блуждания на графе G состоит в том, что на каждом шаге процесса блуждающий объект находится в вершине и перемещается в другую вершину, выбранную случайным равновероятным образом из соседних вершин. Последовательность посещенных вершин является марковской цепью, состояния которой являются вершинами графа.

Авторы интерпретируют задачу выделения сообществ в графе, как задачу кодирования пути, который пройдет блуждатель, и стараются минимизировать длину кода. Поскольку каждое сообщество имеет свой особый бинарный код, то в сообществе каждая из вершин так же имеет свой уникальный внутренний код (вершины из разных сообществ могут иметь одинаковый код). Причем есть дополнительный код выхода из сообщества, который не совпадает с кодами вершин в данном сообществе.

Кодирование пути представляет собой следующий порядок действий: после попадания в сообщество записывается его код, а также внутренний код вершины в которую попал объект. Затем при перемещении внутри одного сообщества фиксируются внутренние кода вершин. При переходе в другое сообщество записывается код выхода из текущего сообщества и код нового.

Время выполнения этого метода приближается к O(n(m + n)).

Walktrap — метод, так же основанный на случайном блуждании. Данный алгоритм основывается на том факте, что короткие случайные блуждания никогда не приводят к выходу из текущего сообщества. На всех вершинах графа некоторым образом вводится метрика и при помощи матрицы вероятностей перехода вершин между сообществами устанавливается, какие вершины следует объединить в единый «кластер». Сложность данного метода в лучшем случае имеет оценку $O(n^2 \log n)$, а в худшем — $O(mn^2)$.

Louvain — алгоритм, который базируется на жадной оптимизации функции модулярности и использует метод «local moving heuristic» (LMH, эвристика локального перемещения).

LMH переносит вершины графа из одного сообщества в другое таким образом, чтобы каждый такой перенос приводил к увеличению значения модулярности. LMH перемещается по вершинам графа в произвольном порядке и заканчивает свою работу после того, как не остается вершин, перемещение которых приводило бы к увеличению модулярности.

Стоит отметить, что алгоритм зависит от порядка перебора вершин в его первой фазе. Эксперименты, проведенные авторами метода, позволяют убедиться в том, что порядок не сильно влияет на результат работы метода, но может в значительной степени влиять на время выполнения, которое в среднем составляет $O(n \log n)$.

За последние годы было разработано несколько улучшений алгоритма Louvain. Эти улучшения направлены либо на повышение эффективности времени, главным образом, за счет использования менее жестких критериев остановки, либо на улучшение качества результата, либо на улучшение времени вычислений путем распараллеливания.

Smart Local Moving — модификация алгоритма Louvain. Оценки времени выполнения авторы метода не приводят, но поскольку это оптимизация алгоритма Louvain, то величина оценки не должна превосходить $O(n \log n)$.

Третий раздел «Оценка качества кластеризации » посвящен описанию различных метрик оценки качества кластеризации.

После работы алгоритма разбиения на сообщества необходимо оценить качество получившегося результата. Для этого используются функции потерь и функционалы качества. Существуют две абсолютно разные ситуации.

В первой ситуации не известно истинное разбиение на сообщества. Подобная ситуация встречается чаще всего, особенно для графов большого размера и реальных данных. В этом случае для оценки качества используется значение функционала модулярности.

Во второй ситуации истинное разбиение известно. Такое возможно в случае графа знакомств друзей пользователя социальной сети, где он самостоятельно распределил всех друзей на сообщества. В этом случае мы можем ввести метрику на разбиениях вершин и посчитать расстояние между истинным и полученным разбиениями. Для таких целей используется аналог редакторского расстояния для разбиений (split-join distance) или нормализованная взаимная информация — информационный критерий для сравнения двух разбиений (NMI).

Четвертый раздел «Практическая часть» состоит из трех основных этапов. Первый этап работы состоит из сравнительного анализа библиотечных алгоритмов LabelPropogation, Infomap, Walktrap, Louvain на различных датасетах. Так же была подсчитана модулярность и NMI, проведено измерение времени выполнения для каждого из алгоритмов кластеризации. Следующий этап заключается в получении собственного датасета с помощью социальной сети «В Контакте» и анализа перечисленных алгоритмов уже на нем. Заключительный этап работы состоит в реализации алгоритма кластеризации Louvain и дальнейшем его тестировании на полученном датасете, полученном в предыдущей части работы. Так же проведено сравнение реализованного алгоритма с библиотечными алгоритмами Louvain и FastGreedy.

ЗАКЛЮЧЕНИЕ

В результате проделанной работы изучен весь необходимый теоретический материал, связанный с алгоритмами кластеризации, так же осуществлена реализация алгоритма Louvain на языке программирования Python и проведен его анализ, рассмотрены оптимизации данного алгоритма. Так же был проведен сравнительный анализ между некоторыми алгоритмами кластеризации на нескольких больших датасетах. Но анализ, который был проведен на «обезличенных» данных, не позволял понять, какой смысл несут в себе получившиеся разбиения. В связи с этим был создан собственный датасет, основанный на данных социальной сети «ВКонтакте», разбит на ground truth сообщества и проанализирован. В результате чего был сформулирован вывод, что некоторые алгоритмы, такие как Walktrap и Louvain, создают «правильное» разбиение, совпадающее по количеству выделенных комьюнити, но содержимое в этих группах — разное.

Основные источники информации:

1. Обзор метаэвристических методов оптимизации Вестн. Самар. Гос. Техн. Ун-та. Сер. Технические науки. 2019, №3(63), -19 стр

2. Hindawi Social Network Community Detection Using Agglomerative Spectral Clustering Volume 2017, Article ID 3719428, 10 стр

3. International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020, Clustering Social Networks using Nature-inspired BAT Algorithm.

4. Bat != Bad: Brief Introduction to The Bat Algorithm [Элекронный ресурс] URL:https://medium.com/it-paragon/bat-algorithm-bad-algorithmb (Дата обращения 29.05.2022).

5. Mathematical Problems in Engineering Volume 2021, Article ID 1485592, 14 стр.

6. Johan Ugander, Brian Karrer, Lars Backstrom, and Cameron Marlow. The anatomy of the facebook social graph. arXiv preprint arXiv:1111.4503, 2011. [Элекронный ресурс] URL: https://arxiv.org/pdf/1111.4503.pdf (Дата обращения 25.04.2022).

7. 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: 10.18500/1816-9791-2017-17-4-441-451.

8. СЛОЖНЫЕ СЕТИ: Введение в теорию / Евин И.А. – Институт машиноведения имени А.А.Благонравова РАН, 2009 г. – 31с