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

Кафедра информатики и программирования

РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ДЛЯ ПОДБОРА ОПТИМАЛЬНОГО АЛГОРИТМА КЛАСТЕРИЗАЦИИ СОЦИАЛЬНЫХ СЕТЕЙ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студентки 2 курса 273 группы
направления 02.04.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Балашовой Татьяны Алексеевны

Научный руководитель:
зав.кафедрой, к.ф.-м.н., доцент _____ Огнева М.В.

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М.В.

подпись, дата

Саратов 2023

ВВЕДЕНИЕ

Актуальность темы.

Данные являются важнейшей частью решения многих задач, которые возникают в различных областях науки и в повседневной жизни. В настоящее время наблюдается значительное увеличение их объемов, а, следовательно, появляется необходимость в разработке новых подходов к обработке и анализу так называемых больших данных.

Одной из важных задач анализа больших данных является задача кластеризации, которая заключается в разбиении множества объектов на сообщества. Данная задача сейчас очень актуальна, поскольку находит широкое применение в различных сферах человеческой жизни (медицина, спорт, музыка, социальная сфера и т.д.) и возникает тогда, когда, например, нужно выделить дружественные сообщества по интересам, разбить страны мира на группы схожих по экономическому положению государств или по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества. Результаты данных разбиений могут помочь совершать прогнозы на будущее.

Одну из лидирующих позиций по производству больших данных занимают в настоящее время социальные сети. В связи с этим задача кластеризации социальных сетей является особенно актуальной.

На данный момент несмотря на существование различных подходов, задача кластеризации социальных сетей в общем случае не имеет однозначного решения. В каждой конкретной ситуации может потребоваться отдельное исследование для выбора подходящего алгоритма, который будет наиболее эффективен. Также открытым остается вопрос оценивания качества кластеризации.

В последнее время также является открытой проблема улучшения точности и ускорения существующих алгоритмов за счет реализации гибридных моделей алгоритмов.

Цель магистерской работы – разработка программного комплекса решения задачи кластеризации социальных сетей, способного в зависимости от исходных данных подбирать наиболее эффективные алгоритмы кластеризации.

Поставленная цель определила **следующие задачи**:

1. Исследовать существующие подходы к решению задачи поиска сообществ в графах, к гибридизации алгоритмов, визуализации данных.
2. Рассмотреть наиболее популярные алгоритмы поиска сообществ в графах, а также существующие метрики оценки качества кластеризации, теоретически выявить достоинства и недостатки алгоритмов.
3. Реализовать разобранные алгоритмы самостоятельно и с помощью методов библиотек, а также алгоритмы для вычисления рассмотренных метрик.
4. Разработать гибридную модель алгоритма, учитывающую особенности реализованных алгоритмов.
5. Разобрать существующие способы классификаций структур сети, подготовить данные, образующие различные сетевые структуры.
6. Провести сравнительный анализ методов на примерах подготовленных данных (тестовых и реальных), визуализировать полученные результаты.
7. Продумать графический интерфейс для пользователя и реализовать комплекс для выбора оптимальных алгоритмов кластеризации социальных сетей.

Методологические основы проблемы поиска сообществ представлены в работах Замятина, Пятницкого, Пастухова [1 – 3].

Практическая значимость бакалаврской работы заключается в реализации метрик качества кластеризации, в работе с не обезличенными, реальными данными социальной сети ВКонтакте, построении сетей

различных структур, а также в реализации программного комплекса кластеризации, содержащего в себе интеграционный сервис для взаимодействия языков программирования C# и Python.

Структура и объём работы. Магистерская работа состоит из введения, 11 разделов, заключения, списка использованных источников и 12 приложений. Общий объем работы – 117 страниц, из них 80 страниц – основное содержание, включая 23 таблицы, список использованных источников информации – 43 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Существующие подходы к поиску сообществ в графах» посвящен проблеме поиска сообществ в графах и способах ее решений в исследованиях различных лет.

Поиск сообществ в графах является сложной задачей в области машинного обучения и обработки данных, поскольку свойства сети могут быть охарактеризованы различными способами – связностью узлов, свойствами связей, семантикой узлов и связей, плотностью графов, а также объемами данных. Кроме того, результат во многом может зависеть от вида сети, получающейся из исходных данных.

В настоящее время существует множество различных подходов к поиску сообществ в графах. Ряд алгоритмов можно разделить в зависимости от механизмов, которые они реализуют. Другие алгоритмы основаны на построении модели графа и вовлечении дополнительных инструментов (нейронные сети, глубокое обучение, стохастическое моделирование, неотрицательная матричная факторизация и т.д.)

Второй раздел «Кластеризация графов: общая постановка задачи» посвящен введению формальной поставки задачи кластеризации графов в целом. Она выглядит следующим образом.

Пусть X – множество объектов, Y – множество номеров кластеров. Задана функция расстояния между объектами $\rho\,(x,\ x')$. Имеется конечная обучающая выборка объектов $X_m = \{x_1, x_2, .., x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров существенно отличались.

Третий раздел «Метрики оценки качества» посвящен описанию различных метрик оценки качества кластеризации – внутренних и внешних.

Внешние методы – методы, основанные на сравнении результата с априори известным разделением на классы, так называемыми, ground-truth сообществами.

Внутренние методы – методы, определяющие качество кластеризации только по признаковым описаниям объектов, без итогового разбиения.

Несмотря на то, что на данный момент существует множество различных методов оценки, проблема оценки качества кластеризации трудноразрешима в силу того, что не существует оптимального алгоритма кластеризации и многие из них не способны определить истинное количество кластеров в данных.

Четвертый раздел «Теоретические описания алгоритмов» посвящен особенностям реализаций рассматриваемых алгоритмов, а также их вычислительной сложности.

В реализации алгоритма Edge Betweenness используется расчет коэффициента центральности «по посредничеству» для каждого ребра с целью оптимизировать модулярность.

Алгоритм Infomap использует механизм случайных блужданий: задача поиска сообществ в графе сводится к минимизации длины кода пути, который проделает «случайный блуждатель».

Алгоритм Labelpropagation основывается на том принципе, что вершина относится к тому сообществу, что и большинство ее соседей.

Целью алгоритма Fastgreedy является жадная оптимизация модулярности.

Алгоритм Walktrap, также как и Infomap, обращается к случайным блужданиям. Основная идея алгоритма заключается в том, что короткие случайные блуждания не приводят к выходу из сообщества.

Алгоритм кластеризации графов Lovain, также, как и Fastgreedy, основан на жадной оптимизации модулярности, однако отличается тем, что зависит от перебора вершин на этапе оптимизации модулярности.

Алгоритм Smart Local Moving является оптимизацией алгоритма Lovain, вычислительно эффективен, способен определять сообщества как в сетях малого и среднего размера, так и в больших сетях с десятками миллионов узлов и с сотнями миллионов ребер.

К достоинствам алгоритма Multilevel можно отнести то, что он интуитивно-понятен и прост в реализации, а также чрезвычайно быстр – исследователи утверждают, что он работает за линейное время для разреженных данных. Также, исходя из описания данного алгоритма, он может работать со взвешенными графами.

Пятый раздел «Структуры социальных сетей» посвящен разбору типов структур сетей, основанных на принципе построения связей между вершинами. По этому критерию все сети, представляющие собой объекты реального мира, можно разделить на три крупные группы:

- граф связей – граф, в котором ребро между вершинами проводится в случае, если между соответствующими им пользователями задано какое-нибудь отношение (например, дружба или родственные связи);
- граф контента – граф, образующий контент, созданный самими пользователями, ребро между вершинами в котором проводится в случае, если, например, два пользователя подписаны на одно сообщество или состоят в одной группе по интересам;
- граф взаимодействий – граф, ребро между вершинами в котором проводится в случае, если между пользователями существует какое-то взаимодействие (например, переписка).

Шестой раздел «Описание модуля для анализа данных социальной сети ВКонтакте» посвящен описанию специального модуля для создания скриптов для социальной сети ВКонтакте – VK API. С его помощью можно полноценно пользоваться функционалом данной социальной сети – отправлять сообщения, просматривать фотографии, списки друзей, аудиозаписей, публиковать записи на стене и т.д.

Для совершения этих манипуляций в данном модуле существуют определенные методы, которые подразделяются в зависимости от объектов сети, с которыми они работают, на следующие категории: аккаунт, друзья, реклама, базы данных, документы, аудиозаписи и т.д.

В данной работе модуль VK API рассматривается как инструмент для осуществления сбора данных с последующей их обработкой и анализом (парсинга).

Седьмой раздел «Реализация алгоритмов кластеризации» посвящен реализации алгоритмов кластеризации, а также алгоритмов для расчета метрик качества. В данном исследовании с помощью средств библиотеки igraph были реализованы такие алгоритмы кластеризации, как InfoMap, Labelpropagation, Fastgreedy, Walktrap, Multilevel; с помощью средств библиотеки NetworkX был реализован алгоритм Lovain; самостоятельно были реализованы алгоритмы Smart Local Moving, а также гибридная модель алгоритмов.

Восьмой раздел «Расчет оценок качества» посвящен реализации алгоритмов для расчета внутренних и внешних методов оценки качества кластеризации.

В данной работе самостоятельно были реализованы алгоритмы для расчета внешних оценок качества кластеризации (индекс Rand, индекс Фоулкса-Мэллова, индекс Жаккара, индекс Phi, Entropy, Purity, F-мера, Variation Of Information) и внутренних (силуэт, индекс BSS, индекс WSS).

Девятый раздел «Построение различных структур социальной сети ВКонтакте» посвящен разработке алгоритмов построения трех структур графов на основании данных социальной сети ВКонтакте.

Особенность реализации методов построения структур графов заключается в том, что какие-то страницы могут оказаться закрытыми, замороженными, заблокированными, удаленными. Как следствие, нельзя получить доступ к информации данной страницы. В таком случае вершина в графе, соответствующая этой странице, будет изолирована.

Десятый раздел «Реализация программного комплекса» посвящен описанию процесса реализации модулей программного комплекса для подбора оптимального алгоритма кластеризации.

В комплексе рассматриваются восемь алгоритмов кластеризации графов – Infomap, Labelpropagation, Fastgreedy, Walktrap, Multilevel, Lovain, Smart Local Moving, а также гибридный алгоритм, реализованный самостоятельно.

Серверная часть приложения реализована на языке программирования Python, клиентская часть приложения – на языке программирования C#.

Одиннадцатый раздел «Сравнительный анализ» посвящен проведению сравнительного анализа различных алгоритмов кластеризации сетей.

В первой подчасти данного раздела анализ проводился на искусственных, обезличенных данных, чье внутреннее содержание, как следствие, нельзя посмотреть.

На данном этапе удалось выделить группы алгоритмов со схожими параметрами и выдвинуть гипотезу о том, что алгоритмы действительно можно разделить в зависимости от реализуемого механизма и на разных структурах сетей они будут работать по-разному. Однако результат может оказаться не гарантированным, поскольку данные «обезличены» и можно получить только верхне-уровневое их описание. Поэтому имеет смысл рассмотреть реальные данные.

Во второй подчасти данного раздела был проведен сравнительный анализ алгоритмов на рассматриваемых структурах сетей самостоятельно-сгенерированного набора данных социальной сети ВКонтакте.

В результате проведенного сравнительного анализа на реальных данных удалось выявить предположение о зависимости применение алгоритмов от структуры сети.

В третьей подчасти данного раздела был произведен подбор алгоритмов с помощью реализованного комплекса, результат был сопоставлен с выдвинутым предположением.

Результаты подбора алгоритмов, полученные с помощью комплекса, во многом схожи с гипотезой, выдвинутой аналитическим способом.

ЗАКЛЮЧЕНИЕ

В рамках данной работы были рассмотрены современные подходы к решению задачи поиска сообществ в графах с различной структурой сети, к гибридизации алгоритмов и визуализации данных.

Были изучены некоторые алгоритмы жесткой кластеризации графов – Edge Betweenness, Infomap, Labelpropagation, Walktrap, Fastgreedy, Lovain, Smart Local Moving, Multilevel. Теоретически были выявлены достоинства и недостатки каждого алгоритма.

На языке программирования Python с помощью библиотеки igraph были реализованы такие алгоритмы кластеризации, как Infomap, Labelpropagation, Fastgreedy, Walktrap, Multilevel; с помощью библиотеки networkx был реализован алгоритм Lovain, алгоритм Smart Local Moving, а также разработана и реализована модель гибридного алгоритма, объединяющего подходы жадной оптимизации модулярности и распространения близости. Были реализованы алгоритмы для вычисления метрик оценки качества кластеризации.

Были разработаны алгоритмы для построения различных видов структур графов, построены графы различных структур, сами же вершины были вручную разбиты на ground-truth сообщества. Изучена внутренняя структура разбиений, получившихся в результате работы алгоритмов, сделаны выводы.

Был разработан метод для осуществления раскраски графа в зависимости от кластера для проведения загрузки графа в платформу визуализации, проведена непосредственная визуализация полученных результатов с помощью платформы Gephi.

На языке программирования Python был реализован GRPC сервер, предоставляющий набор алгоритмов машинного обучения и теории графов, с помощью специального интеграционного соглашения между языками программирования Python и C#.

На языке программирования C# было реализовано трехслойное приложение – комплекс, позволяющий в зависимости от исходных данных подбирать оптимальные разбиения.

Было организовано распараллеливание алгоритмов построения графов различных структур сетей с помощью многопоточности, асинхронного выполнения, а также библиотеки Task Parallel Library, проведено сравнение времени выполнения каждого подхода, сделаны выводы.

На основании всех рассмотренных материалов и результатов, полученных на практике, можно сделать вывод о том, что задача поиска сообществ в графах до сих пор является нетривиальной, актуальной, активно изучаемой задачей. Структура сети является очень важной составляющей при анализе результатов. Отдельно стоит отметить про механизм оценки качества – алгоритм может выигрывать в одной метрике, но проигрывать в другой. Поэтому нужно вводить приоритет метрик, что непосредственно и было сделано в рамках реализованного комплекса.

Результаты работы были опубликованы и представлены на следующих конференциях:

- XIII научно-практическая конференция «Presenting Academic Achievements to the World».
- XIV всероссийская научно-практическая конференция «Информационные технологии в образовании».
- итоговая студенческая научная конференция СГУ.

Основные источники информации:

1. Замятин, В. М. Обзор методов кластеризации в системах социологического тестирования / В. М. Замятин // Инновационно-инвестиционный фундамент развития экономики общества и государства: от научных разработок к практике. – 2021. – С. 33– 38.
2. Пятницкий, А. М. Поиск статистически значимых кластеров в сетях и медикобиологические приложения / А. М. Пятницкий, В. М. Гукасов // Медицина и высокие технологии. – 2022. – № 1. – С. 53-62.
3. Пастухов, Р. К. Определение влиятельных пользователей социальной сети по двудольному графу комментариев / Р. К. Пастухов, М. Д. Дробышевский, Д. Ю. Турдаков // Труды Института системного программирования РАН. – 2022. – Т. 34, № 5. – С. 127-142.
4. Шестаков, Т. А. Интеллектуальный анализ информации о пользователях социальных сетей / Т. А. Шестаков, Ю. А. Леонов, А. А. Кузьменко, А. С. Сазонова, Р. А. Филиппов // Прикладная математика и вопросы управления. – 2021. – № 4. – С. 72-91.
5. Dhumal, A. Survey on Community Detection in Online Social Networks / A.Dhumal, P.Kamde // International Journal of Computer Applications. - 2015. – Т. 121, № 9. – P. 35–41.
6. Маковский, В. Н. Модель системы связи с динамической топологией сети на основе многоэлементных регулярных кластерных структур в информационном поле представления дискретных систем / В. Н. Маковский, А. М. Рахматулин, А. Н. Смирнов // Журнал радиоэлектроники. – 2021. – № 4.
7. Жукова, Н. А. О проблеме сбора данных в сетях интернета вещей с динамической структурой (обзор) / Н. А. Жукова, А. Б. Тристанов, Т. Тин, М. Аунг // Известия КГТУ. – 2021. – № 61. – С. 105-118.