МИНОБРНАУКИ РОССИИ

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

Кафедра социальной информатики

КЛАСТЕРИЗАЦИЯ ИНТЕРНЕТ СООБЩЕСТВ

(автореферат бакалаврской работы)

Студента 5 курса 531 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Микаелян Микаел Мартуновича

Научный руководитель
кандидат физико-математических наук,
доцент ________ Л.Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2021

ВВЕДЕНИЕ

Актуальность проблемы. В современном мире, где технологии развиваются особенно быстро, растет и объем производимой и потребляемой информации. Это создает необходимость совершенствовать старые способы ее обработки, но и изобретать новые, отвечающие требованиям времени, так как появляется всё большее количество данных, которые необходимо хранить и обрабатывать для безопасной и удобной работы с информацией. Качественное хранение, обработка и представление данных стали как никогда востребованными. Развитие информационных технологий повлияло на значительное упрощение взаимодействия между людьми. Каждодневно любой из нас общается с широким количеством людей, прямо или косвенно.

Коммуникации между людьми формируют своеобразную социальную сеть. Термин «социальная сеть» впервые был введен социологом Джеймсом Барнсом: «социальная сеть – это социальная структура, состоящая из группы узлов, которыми являются социальные объекты (люди или организации), и связей между ними (социальных взаимоотношений)». В настоящее время под этим понятием почти всегда понимается платформа в информационно-телекоммуникационной сети «Интернет», хотя алгоритмы, применимые к такого рода сетям не теряют свою актуальность и при анализе тех, которые не связаны с интернетом.

В этом случае социальная сеть может предоставить довольно большое количество данных о своих пользователях, что может упростить процесс кластеризации и, вместе с тем, повысить его точность и эффективность. Анализ социальных сетей дает возможность многое узнать о характеристиках ее элементов, а также об их взаимодействии с другими элементами этой сети. Для кластерного анализа ее необходимо представить в виде графа.

Актуальность темы бакалаврской работы обусловлена тем, что некоторые компании могут использовать кластеризацию графа социальной сети для выведения кластеров клиентов и их характеристик, чтобы четко понимать, чего хочет потребитель, какие услуги или товары были бы ему полезны. Это один из

наиболее эффективных способов получить информацию от клиента, поскольку для этого требуется лишь определенная вычислительная мощность и относительно небольшие затраты времени. Чтобы провести кластерный анализ, потребуется грамотный отбор содержательных входных данных, которые необходимы для получения необходимого результата. Когда речь заходит о кластеризации именно социальной сети, в набор этих данных обязательно включаются связи между ее объектами (узлами). На выходе получается информация о кластерах клиентов, обладающих схожими характеристиками и сравнительно более тесными связями между объектами.

Степень научной разработанности данной проблемы: Проблемы кластеризации интернет сообществ рассматривались в трудах И. Д. Манделя [1], С. А. Суслов [2], С.В Мачульскис, [3], К.В Воронцов [4], В области Модификация методов анализа социальных графов на основе применения атрибутивных компонентов учетных записей для идентификации сообществ пользователей социальных сетей, такие как работы ученых: Синадский, Н.И., Сушков, П.В.

Особое место занимают работы В. С. Бериков,[5] и Г. С. Лбов [6], в которых всесторонне рассмотрены современные тенденции в кластерном анализе.

Целью бакалаврской работы проведение анализа социально сети с использованием графов.

---

[1] Мандель, И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статика, 2019. – 176 с.

[2] . Суслов, С. А. Кластерный анализ: сущность, преимущества и недостатки/ С. А. Суслов. // Вестник НГИЭИ. – 2016. – №1. –. 51-56. С.

[3] Мачульскис, С.В. Исследование многоуровневых алгоритмов кластеризации для визуализации графов большого объема // Материалы 51-й международной научной студенческой конференции «Студент и научно-технический прогресс». Информационные технологии/. Новосибирск, 2016 г. - 138 с.

[4] Воронцов, К.В. Методы кластеризации: курс лекций. Режим доступа: URL: http://www.machinelearning.ru/wiki/(дата обращения 20.05.21)

[5] Бериков, В. С. Современные тенденции в кластерном анализе / В. С. Бериков, Г. С. Лбов. – Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению "Информационно-телекоммуникационные системы", 2019. – 26 с

[6] Бериков, В. С. Современные тенденции в кластерном анализе / В. С. Бериков, Г. С. Лбов. – Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению "Информационно-телекоммуникационные системы", 2019. – 26 с

Задачи работы:

1. Просмотреть всю имеющуюся литературу по тематике, сделав выборку соответствующих материалов, подчеркивающих и отражающих работу моделей и методов кластерного анализа.
2. Изучить существующие в настоящее время методы кластерного анализа.
3. Провести тщательный анализ, изучить эффективность действия самых популярных методов.
4. Указать на преимущества и недостатки при применении популярных методов.
5. Проанализировать сеть с помощью графа

Структура выпускной квалификационной работы представлена введением, четырьмя разделами, заключением и списком использованных источников.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Актуальные проблемы кластерного анализа» дана основная характеристика проблем кластерного анализа таких как:

1. Обоснование качества результатов. Проблема заключается в том, что один и тот же объект может быть классифицирован в различные группы вне зависимости от его внутренних свойств, а в связи с различными экспертными данными или различным построением системы. Для избегания этого необходимо разрабатывать и вводить актуальные критерии качества.

2. Анализ большого числа разнотипных данных порождает методологическую проблему выбора метрик. Также увеличение числа объектов даже однотипных данных может повлечь за собой неразличимость расстояний.

3. Нелинейность взаимосвязей. Классические методы снижения размерности в кластерном анализе направлены на линейной взаимосвязи между переменными. Для поиска более сложных зависимостей необходимо переходить к ядерным методам.

Немаловажной здесь является проблема поиска глобального экстремума функции критерия качества. Критерий качества, как правило, является

функцией, зависящей от большого числа факторов, нелинейной, обладающей множеством локальных экстремумов. Для нахождения кластеров необходимо решить сложную комбинаторную задачу поиска оптимального варианта классификации, поэтому алгоритм полного перебора вариантов – процесс достаточно трудоемкий, экспоненциально зависящий от размерности, особенно если количество групп заранее не известно. Поэтому когда размер таблиц данных увеличивается, происходит «комбинаторный взрыв».

Классические алгоритмы кластерного анализа осуществляют направленный поиск в сравнительно небольшом подмножестве пространства решений, используя различного рода априорные ограничения (на число кластеров или их форму, на порядок включения объектов в группы и т.д.). При этом нахождение строго-оптимального решения не гарантируется. Для поиска оптимального решения применяются более сложные методы, такие как генетические (эволюционные) алгоритмы, нейронные сети и т.д. Существуют экспериментальные исследования, подтверждающие преимущества таких алгоритмов перед классическими алгоритмами. Однако и при использовании эволюционных методов возникают проблемы, связанные со спецификой решаемой задачи кластер-анализа: с трудностью интерпретации используемых операторов рекомбинации и кроссовера.

4. Неустойчивость результатов кластеризации. Зачастую результаты группировки могут сильно меняться в зависимости от множества факторов: выбора начальных условий, порядка объектов, параметров работы алгоритмов и т.д. Многими учеными уже предлагаются способы повышения устойчивости группировочных решений, основанные на применении ансамблей алгоритмов. При этом используются результаты группировки, полученные различными алгоритмами, или одним алгоритмом, но с разными параметрами настройки, по различным подсистемам переменных и т.д. После построения ансамбля проводится нахождение итогового коллективного решения.

5. Недостаточность знаний об объекте. Существует проблема трудноформализуемых областей, в которых становится затруднительным

создание модели объекта и применение алгоритмов, основывающихся на представлении класса, как набора распределённых в пространстве переменных.

6. Проблема представления результатов. Помимо хорошей прогнозирующей способности для любого алгоритма анализа данных важно, насколько понятными и интерпретируемыми являются его результаты. Для улучшения интерпретируемости решений можно использовать логические модели. Такие модели используются для решения задач распознавания образов и прогнозирования количественных показателей, например, в методах построения решающих деревьев или логических решающих функций.

Вторая глава работы «Анализ методов кластерного анализа» посвящен анализу методов кластерного анализа, их разновидностям, сравнение их между собой, а также выделение достоинств и недостатков каждого из них.

Кластерный анализ или кластеризация – это задача группировки набора объектов таким образом, что объекты в одной и той же группе (называемой кластером) были более схожи (в том или ином смысле) друг с другом, чем с другими группами (кластерами).

Кластерный анализ – это не конкретный алгоритм, а общая проблема, которую необходимо решить. Это может быть достигнуто с помощью различных алгоритмов, которые существенно различаются как по своему пониманию того, что такое кластер, так и по подходам к его эффективному поиску. Популярными понятиями кластеров являются группы с небольшими расстояниями между членами кластера, плотные области пространства данных, интервалы или отдельные статистические распределения. Поэтому кластеризация может быть сформулирована как задача многоцелевой оптимизации. Соответствующий алгоритм кластеризации и настройки параметров (включая такие значения, как используемая функция расстояния, порог плотности или количество ожидаемых кластеров) зависят от индивидуального набора данных и предполагаемого использования результатов.

Кластерный анализ как таковой является итеративным процессом обнаружения знаний или интерактивной многоцелевой оптимизации, которая включает в себя метод проб и ошибок.

Зачастую необходимо работать и видоизменять именно предварительную обработку данных и параметры модели до тех пор, пока не будет достигнут желаемый результат.

В третьей главе «Модели кластерного анализа» описаны несколько моделей кластерного анализа. Первой рассмотренной моделью будет модель Барабаши — Альберт, которая позволяет синтезировать случайные безмасштабные сети с использованием принципа предпочтительного присоединения. Принцип предпочтительного присоединения заключается в том, что в каждый момент времени в синтезируемую сеть добавляется новый узел, который соединяется с существующими узлами с вероятностью, пропорциональной числу связей этих узлов. Синтезируемая социальная сеть является графом, узлы которого представляют социальные объекты (пользователей), а ребра — социальные связи между ними. редложенная модификация алгоритма Барабаши — Альберт используется при формировании массивов данных, имитирующих взаимодействие в социальных сетях, с учетом заданного шаблона взаимодействия пользователей.

Второй рассмотренной моделью была триадная модель. Многоагентная триадная модель сети представляет собой множество взаимосвязанных агентов (субъектов или объектов), в котором:

1. Каждый агент сохраняет свою индивидуальность, а именно, имеет собственные (индивидуальные) цели, выполняет направленные на достижение этих целей индивидуальные действия, характеризуется индивидуальными показателями;
2. Связанность агентов заключается в том, что их деятельность может координироваться во времени, и в определенные моменты они могут передавать друг другу ресурсы;

3. Результатом индивидуальной деятельности агентов является достижение определенных коллективных целей и определенная динамика коллективных показателей.

Задача синтеза сети в общем виде формулируется следующим образом. Известен набор агентов, представленных, например, расширенными триадными структурами, отображающими взаимовлияние графов индивидуальных целей, действий, показателей, и заданы коллективные цели и показатели. Требуется определить, можно ли путем организации связей между агентами создать сеть, в которой наряду с индивидуальными целями достигались бы и желаемые коллективные цели при допустимых значениях индивидуальных и коллективных показателей.

Третьей рассматриваемой моделью была сетевая модель. Сетевая модель, также относящаяся к теоретико-графовым, явилась развитием подходов, реализованных в модели иерархической. Прежде всего, развитие касается связей между записями, имеющих двунаправленный характер. Сетевую модель можно представить в виде графа с узлами в виде записей, и рёбрами, отображающими наборы. Направление и характер связи в сетевых БД не являются очевидными, как в иерархических БД, поэтому характеристики и направление связей должны указываться явно при описании БД.

В 1975 году конференция CODASYL (Conference of Data System Languages) стандартизовала базовые понятия и формальный язык описания. Широко известной системой, основанной на сетевой модели данных, являлась СУБД IDMS (Integrated Database Management System), использовавшаяся на мэйнфреймах IBM. В настоящее время IDMS принадлежит компании Computer Associates, имеющей неформальный статус «лавки старьёвщика» в мире софтостроения: как правило, все купленные компанией продукты берутся на сопровождение, но не развиваются, доживая до своего естественного конца.

Несмотря на некоторые интересные особенности и преимущества, до наших дней дожило только небольшое число СУБД, реализующих сетевую модель, например американская Raima (бывшая dbVista) и отечественная

КроносПро. На их примере также можно проследить эволюцию программных продуктов класса сетевых СУБД.

СУБД Raima изначально использовалась, как встроенная с достаточно низкоуровневым API для языка Си. Постепенно, в систему был добавлен интерфейс доступа на SQL, а сама СУБД получила возможность работы в режиме клиент-сервер. По-прежнему Raima используется для транзакционных приложений как лёгкое (lightweight) кросс-платформенное решение.

Напротив, развитие СУБД Кронос пошло в сторону аналитической обработки. Это та область, где преимущества сетевой модели могут быть использованы полнее, а недостатки обойдены более просто. Для объяснений нам все же придётся кратко познакомиться с основными понятиями сетевой модели.

Термин запись соответствует аналогичному понятию структурного типа в традиционных языках программирования: record в Паскаль-подобных или struct в наследниках Си.

Было отмечено, что одной из особенностей аналитической обработки является нахождение базы данных в режиме чтения, изменения связей между записями редки и происходят, как правило, только в моменты обновления информации. Поэтому использование сетевой модели в аналитических приложениях может нивелировать влияние указанных недостатков.

В четвертой главе «Анализ социальной сети с использованием графов» посвящен анализу социальной сети с использованием графов. В этом разделе мы подробно проанализировали работу сети и получили некоторый результат описанный в ВКР. Анализ социальной сети – это процесс исследования разных по свойствам систем с использованием теории сетей. Он начал обширно применяться именно тогда, когда стало ясно, что большое кол-во существующих социальных, экономических и биологических сетей обладают уникальными свойствами: изучив один тип, можно понять структуру и любых других сетей и научиться делать предсказания по ним.

Любые сети состоят из отдельных участников (людей или вещей в сети) и отношений между ними. Сети очень часто визуализируются с помощью графов – структур, состоящих из множества точек и линий, отображающих связи между этими точками. Участники представлены в виде узлов сети, а их отношения представлены в виде линий, которые связывают их между собой.

С помощью анализа социальных сетей можно проанализировать самые разные взаимодействия и процессы обмена разными материальными и информационными ресурсами. К примеру, если проанализировать сеть транзакций между клиентами банка (где каждый узел будет являться банковским клиентом, а рёбрами – переводы между ними), можно определить круг лиц, которые вовлечены в мошеннические операции, или можно так же выявить нарушения внутренних регламентов сотрудниками банка.

Также можно выстроить сеть рабочих взаимоотношений (на примере разных типов коммуникаций среди сотрудников), которая поможет в понимании социальной структуры организации и позиции любого работника в этой структуре. При наличии данных о типе коммуникации для каждого работника возможно даже проанализировать, как такие характеристики как лидерство, наставничество и сотрудничество воздействуют на его карьеру, а на основании полученных знаний установить карьерные цели и предложить тренировочные программы для их достижения.

Кроме того, на примере сетей можно и прогнозировать события. Например, имеются модели, которые оценивают вероятность сбоя программного обеспечения, некоторые из них рассматривают людей как источник для прогнозов – ведь именно люди разрабатывают и тестируют продукты до релиза. При их взаимодействии образуется сеть: можно представить каждого сотрудника как узлы, а то, работали ли они вместе над одним и тем же файлом в рамках одного релиза, как рёбра сети. Понимание взаимодействий и информация о ранее произошедших сбоях позволит многое сказать о надёжности конечного продукта и укажет на файлы, в которых риск сбоя наиболее вероятен.

ЗАКЛЮЧЕНИЕ

В результате дипломного проектирования были рассмотрены методы кластерного анализа, популярные продукты для кластеризации данных и их применение. На основании чего было установлено, что множество систем реализованы под конкретные задачи. Для решения задачи кластеризации были изучены и опробованы графовые и статистические методы. Графовая кластеризация более удобна для визуализации итога проведенной работы.

В ходе выполнения ВКР достигнуты следующие результаты:

1. Сделана выборка соответствующих материалов из литературы по заданной тематике, которая подчёркивает и отражает работу моделей и методов кластерного анализа.
2. Изучены существующие в настоящее время методы кластерного анализа.
3. Изучена эффективность действия самых популярных методов.
4. Указаны преимущества и недостатки при применении популярных методов.
5. Проанализирована сеть с помощью графа