**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«Сибирский государственный университет науки и технологий**

**имени академика М.Ф. Решетнева»**

ИИТК

институт

Кафедра системного анализа и исследования операций

кафедра

**ДНЕВНИК-ОТЧЕТ**

**ПРОХОЖДЕНИЯ ПРАКТИКИ ОБУЧАЮЩИМСЯ**

Калинин

(фамилия)

Дмитрий Александрович

(имя, отчество)

Направление / специальность 27.03.03 Системный анализ и управление

Группа БСА21-01

Вид практики учебная

Тип практики ознакомительная практика

Сроки прохождения практики с 05.07.2022 по 18.07.2022

Оценка кафедры по результатам прохождения практики \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель от Университета О. В. Чубарова

инициалы, фамилия подпись дата

Красноярск 2022 г

**ПАМЯТКА ОБУЧАЮЩЕГОСЯ**

**При оформлении на практику обучающийся обязан иметь следующие документы:**

* паспорт;
* трудовую книжку, за исключением случаев, когда трудовой договор заключается впервые;
* страховое свидетельство обязательного пенсионного страхования (СНИЛС);
* документы воинского учета – для военнообязанных и лиц, подлежащих призыву на военную службу;
* идентификационный номер налогоплательщика (ИНН);
* предписание и справку-форму для обучающихся, проходящих практику в режимных Профильных организациях;
* методические указания по организации практики;
* направление от Университета в Профильную организацию;
* настоящий дневник-отчет с заполненными разделами.

**В период прохождения практики обучающийся обязан:**

* прибыть на место прохождения практики в сроки, установленные календарным учебным графиком;
* выполнить индивидуальное задание, выданное руководителем от Университета;
* соблюдать требования охраны труда и пожарной безопасности, правил внутреннего трудового распорядка, техники безопасности и санитарно-эпидемиологических правил и гигиенических нормативов;
* вести дневник-отчет практики (для обучающихся по программам бакалавриата, специалитета и магистратуры), где отражать ход выполнения индивидуального задания, описывать выполненную работу, и оформить полученные результаты в соответствии с требованиями, установленными программой практики и методическими указаниями.

**По окончании практики обучающийся обязан:**

* сдать на предприятие всю документацию, которой он пользовался в период практики;
* получить справку-форму о допуске в режимную Профильную организацию для возврата её в 1-й отдел Университета (для режимной Профильной организации);
* сдать пропуск в Профильную организацию;
* своевременно оформить и сдать дневник-отчет по итогам практики руководителю от Университета.

**Обучающемуся необходимо знать:**

* при подведении итогов работы обучающегося принимается во внимание оценка результатов прохождения практики, данная руководителем от Профильной организации, качество дневника-отчета;
* неудовлетворительные результаты промежуточной аттестации по практике или не прохождение промежуточной аттестации по практике при отсутствии уважительных причин признаются академической задолженностью;
* обучающиеся, не ликвидировавшие в установленные сроки академическую задолженность, отчисляются из Университета приказом проректора по образовательной деятельности по представлению директора института как не выполнившие обязанности по добросовестному освоению образовательной программы и выполнению учебного плана.

Ознакомлен\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(подпись, дата)

**Наименование Профильной организации, в которой обучающийся проходит практику[[1]](#footnote-1):**

Научно-исследовательская лаборатория ИИТК

**Руководителем от Университета назначен:**

Чубарова Олеся Викторовна

(фамилия, имя, отчество)

доцент

(должность на кафедре)

Контактный телефон \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Руководителем от Профильной организации назначен[[2]](#footnote-2):**

Попов Алексей Михайлович

(фамилия, имя, отчество)

директор ИИТК, зав. НИЛ ИИТК

(должность в цехе, отделе)

Контактный телефон \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата фактического прибытия обучающегося в

Профильную организацию 05.07.2022

**М. П.**[[3]](#footnote-3)

Дата фактического убытия обучающегося из

Профильной организации 18.07.2022

**М. П.**

Вводный инструктаж провел[[4]](#footnote-4): «\_\_»\_\_\_\_\_\_\_\_\_20\_\_г.

Руководитель службы охраны труда

(должность)

Дубовик Тамара Петровна

(ФИО, подпись)

Инструктаж на рабочем месте провел[[5]](#footnote-5): «\_\_»\_\_\_\_\_\_\_\_\_20\_\_г.

директор ИИТК, зав. НИЛ ИИТК

(должность)

Попов Алексей Михайлович

(ФИО, подпись)

**Индивидуальное задание на практику**

|  |  |
| --- | --- |
| **№ п/п** | **Наименование работ** |
| 1 | Разработать программу, реализующую алгоритм автоматической группировки (кластеризации) k-средних. |
| 2 | Разработать программу, реализующую алгоритм автоматической группировки (кластеризации) k-средних c манхэттенской (прямоугольной) мерой расстояния. |
| 3 | Оформление отчета |
| 4 | Презентация работы |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Задание выдал: Руководитель от Университета

Чубарова Олеся Викторовна

(Ф.И.О., подпись, дата)

Задание согласовал: Руководитель от Профильной организации

Попов Алексей Михайлович

(Ф.И.О., подпись, дата)

**Рабочий график проведения практики**

|  |  |  |
| --- | --- | --- |
| **Период** | **Краткое описание выполняемых работ** | **Отметка руководителя от Профильной организации о качестве выполнения работ** (отлично/ хорошо/ удовлетворительно) |
| 08.07.2022- 11.07.2022 | Разработать программу, реализующую алгоритм автоматической группировки (кластеризации) k-средних. |  |
| 11.07.2022- 13.07.2022 | Разработать программу, реализующую алгоритм автоматической группировки (кластеризации) k-средних c манхэттенской (прямоугольной) мерой расстояния. |  |
| 13.07.2022 - 18.07.2022 | Оформить отчет-дневник |  |
| 19.07.2022 | Презентовать проделанную работу |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Руководитель от Профильной организации

Попов Алексей Михайлович

(Ф.И.О., подпись, дата)

**Отчет о прохождении учебной практики**

(*вид практики*)

**Примерное содержание отчета[[6]](#footnote-6)**

1. Введение

2. Общая часть

3. Заключение (Выводы)

4. Библиографический список

5. Приложения (при необходимости)

**Введение**

Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на подгруппы, именуемые кластерами. Внутри каждого такого кластера должны оказаться объекты, схожие между собой по различным признакам, а объекты разных групп, в свою очередь, должны быть как можно более отличны[1].

Также стоит выделить отличия между кластеризацией и классификацией. Состоит это отличие в том, что перечень кластеров зачастую не определён в изначальных условиях задачи, неизвестно количество кластеров. Эти показатели определяются в процессе работы алгоритма[1].

Ещё одно отличие между задачами классификации и задачами кластеризации состоит в том, что решение задач кластеризации являет собой входит в прикладную область решения задач машинного обучения без учителя, в то время как задачи классификации входят в область задач обучения с учителем.

Отношение задач классификации к области задач машинного обучения без учителя также накладывает свои ограничения на методы, применяемые к решению задач такого типа, увеличивает сложность их решения.

Одним из популярных алгоритмов решения задачи кластеризации является метод k-средних[4]. Однако, этот алгоритм не всегда может адекватно решить поставленную задачу[1].

В данной работе представлена реализация другого алгоритма решения задач кластеризации – метода k-medoids. Этот алгоритм реализуется для выполнения операций над предварительно подготовленными данными, полученными из текста, в котором попарно было найдено Жаккардово расстояние между различными строками.

**Предлагаемая методика**

Кластерный анализ выполняет следующие основные задачи:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Порождение гипотез на основе исследования данных.
4. Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

1. Отбор выборки для кластеризации.
2. Определение множества переменных, по которым будут оцениваться объекты в выборке.
3. Вычисление значений той или иной меры сходства между объектами.
4. Применение метода кластерного анализа для создания групп сходных объектов.
5. Проверка достоверности результатов кластерного решения.

Кластерный анализ предъявляет следующие требования к данным: во-первых, показатели не должны коррелировать между собой; во-вторых, показатели должны быть безразмерными; в-третьих, их распределение должно быть близко к нормальному; в-четвёртых, показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов; в-пятых, выборка должна быть однородна, не содержать «выбросов».

Цели кластеризации имеют все шансы быть разными, однако среди их числа можно выделить 3 главных направления.

Первое – это данное представление информации – подразумевает разделение подборки в категории сходных предметов, то, что дает возможность облегчить последующее обрабатывание информации и реализовать заключение, применяя к любому кластеру свой собственный способ рассмотрения. В качестве образца возможно просмотреть биосистематику – научную дисциплину, в задачи которой входит создание основ систематизации живых существ фактическое дополнение данных принципов к построению системы.

Второе течение – сокращение информации – предназначается для сокращения размеров информации с наименьшими утратами данных. В случае если начальная выборка чрезмерно огромна, то возможно уменьшить ее, оставив по одному наиболее обычному представителю от любого кластера. Нередко рассмотрение единственного предмета может быть достаточной ради нахождения черт абсолютно всех сходных объектов с значительной возможностью. К примеру, доказав теорему о площади любого прямоугольника, мы можем так же отметить, что она правдива для всех прямоугольников на плоскости.

Третье течение – выявление новизны – направлено на выявление новейших или мало встречающихся объектов в некоторых выборках. При этом выделяются нетипичные объекты, которые никак не получается добавить ни к одному из кластеров. Кластерный анализ имеет ряд достоинств перед другими методами классификации данных. В первую очередь, это связано с тем, что он дает возможность осуществлять разделение объектов не по одному, а по целому комплекту свойств. К тому же, давление любого из характеристик может быть достаточно просто усилено или ослаблено путем внесения в математические формулы определенных коэффициентов.

Помимо этого, кластерный анализ не накладывает ограничений на вид группируемых объектов, и дает возможность рассматривать множество исходных данных практически произвольной природы. Еще одной характерной чертой кластеризации считается то, что многие алгоритмы способны самостоятельно определить число кластеров, на которое следует разбить данные, а также отметить характеристики этих кластеров без участия человека только при помощи применяемого метода.

**Описание основных методов кластерного анализа**

Общепринятой классификации методов кластеризации не существует[2], но можно выделить ряд групп подходов (некоторые методы можно отнести сразу к нескольким группам и потому предлагается рассматривать данную типизацию как некоторое приближение к реальной классификации методов кластеризации)

* Метод k-средних. Наиболее популярный метод кластеризации. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров.
* Метод k-medians. Это вариация k-средних, в котором для определения центроида кластера вместо среднего вычисляется медиана.
* EM-алгоритм. Алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных.
* Алгоритмы семейства FOREL. Алгоритм кластеризации, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения.
* Метод нечеткой кластеризации C-средних. позволяет разбить имеющееся множество векторов (точек) мощностью p на заданное число нечетких множеств.
* Нейронная сеть Кохонена. Класс нейронных сетей, основным элементом которых является слой Кохонена.
* Генетический алгоритм. Это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, аналогичных естественному отбору в природе.
* Иерархическая кластеризация. Совокупность алгоритмов упорядочивания данных, визуализация которых обеспечивается с помощью графов.

**Алгоритм k-medoids**

Алгоритм k-medoids представляет собой вариант популярного k-means[3], однако, в отличие от k-means, в k-medoids в качестве центроидов может выступать не любая точка, а только какие-то из имеющихся наблюдений. Так как в графе между вершинами расстояние определить можно, k-medoids годится для кластеризации графа. Главная проблема этого метода — необходимость явного задания числа кластеров, то есть это не выделение сообществ (сommunity detection), а оптимальное разбиение на заданное количество частей (graph partitioning)[4].

**Алгоритм PAM**

Самый распространённый вариант реализации k-medoids называется PAM (Partitioning Around Medoids). Он представляет собой «жадный» алгоритм с очень неэффективной эвристикой.

За каждую итерацию алгоритм перебирает (n-k) точек графа (где n— общее количество узлов), перемещая туда соответствующий медоид. Для каждой такой замены ему нужно пересчитать расстояния всех точек до медоидов. Если все попарные расстояния между точками влезают в память, то этот этап займет (n – k) \* k действий. Оптимальная замена занимает еще (n – k) \* k действий. Количество итераций для графа с несколькими тысячами узлов — порядка 30-50[4].

**Метрика Жаккара**

Сходство Жаккара — это простая, но иногда мощная метрика сходства. Даны две последовательности A и B: находим число общих элементов в них и делим найденное число на количество элементов обеих последовательностей[5].

Соответственно, расстояние в Жаккардовой метрике вычисляется как разность между 1 и, собственно, сходством Жаккара.

Для реализации алгоритма выражения попарных расстояний между двумя строками текста эти самые строки необходимо представить в виде множеств, в которых нет элементов, которые повторяются.

Затем реализуется сам метод получения расстояния Жаккара – нахождение элементов, принадлежащих обоим множествам, нахождение общего числа элементов в обоих множествах, вычисление частного между количеством первых и вторых, и вычитание этого числа из единицы.

В результате таких вычислений получается матрица, похожая на представленную в таблице 1. В ней выражены попарные Жаккардовы расстояния между различными строками текста.

Таблица 1 – Матрица попарных расстояний в метрике Жаккара.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0.0 | 0.666 | 0.6 | 0.833 | 1.0 | 0.833 | 0.666 |
| 0.666 | 0.0 | 0.6 | 0.833 | 0.875 | 1.0 | 0.857 |
| 0.6 | 0.6 | 0.0 | 1.0 | 1.0 | 0.8 | 0.833 |
| 0.833 | 0.833 | 1.0 | 0.0 | 1.0 | 1.0 | 0.833 |
| 1.0 | 0.875 | 1.0 | 1.0 | 0.0 | 1.0 | 1.0 |
| 0.833 | 1.0 | 0.8 | 1.0 | 1.0 | 0.0 | 0.6 |
| 0.666 | 0.857 | 0.833 | 0.833 | 1.0 | 0.6 | 0.0 |

Как можно заметить, матрица симметрична относительно главной диагонали. Сама же по себе главная диагональ заполнена нулями, по понятной причине: строка максимально схожа сама с собой, а значит и Жаккардово расстояние от строки до самой себя будет равно нулю.

К приведённой таблице был применён описанный ранее алгоритм PAM, поделивший всю таблицу на 2 кластера.

Сама же Жаккардова метрика может в том числе применяться для решения задач сравнения социального окружения[6].

**Заключение**

В ходе работы были рассмотрены основные положения кластерного анализа, приведены и рассмотрены некоторые метрики для изучения текста, рассмотрены основные алгоритмы кластерного анализы, проанализирован и воплощен в программном виде метод реализации одного из этих алгоритмов на реальных данных.

Была произведена кластеризация реальных данных алгоритмом PAM, подтверждена мысль о том, что алгоритм этот жадный до ресурсов, и выполняется чрезвычайно медленно при больших объёмах данных.

Рассмотрены различные варианты предобработки текстов при помощи высокоуровневых языков программирования. Были построены модели, основанные на полученных данных и имеющихся наблюдениях, приведены математические действия, связанные с этим, произведена кластеризация обработанных данных.

**Библиографический список**

1. Обзор алгоритмов кластеризации данных [Электронный ресурс] URL: https://habr.com/ru/post/101338/
2. Классификация методов региональной кластеризации [Электронный ресурс] URL: https://cyberleninka.ru/article/n/ klassifikatsiya-metodov-regionalnoy-klasterizatsii/
3. Understanding K-Means, K-Means++ and, K-Medoids Clustering Algorithms [Электронный ресурс] URL: https://towardsdatascience.com/understanding-k-means-k-means-and-k-medoids-clustering-algorithms-ad9c9fbf47ca
4. Кластеризация графов и поиск сообществ. Часть 2: k-medoids и модификации [Электронный ресурс] URL: https://habr.com/ru/ company/dca/blog/264811/?ysclid=l5qfnhgu5r406242491
5. Документация sklearn.metrics.jaccard\_score [Электронный ресурс] URL: https://scikit-learn.org/stable/modules/generated/ sklearn.metrics.jaccard\_score.html
6. Применимость коэффициентов сходства в задаче сравнения социального окружения [Электронный ресурс] URL: https://scm.etu.ru/assets/files/2020/scm20/papers/1/039.pdf

**Отзыв руководителя от Профильной организации**

**о прохождении практики обучающимся**

1. Полученные компетенции в соответствии с рабочей программой практик

|  |
| --- |
| В ходе учебной практики Калинин Д.А. приобрел способность разработки |
| алгоритмов для решения исследовательских задач в области системного |
| анализа и управления. Усовершенствовал свои навыки разработки |
| программного кода и проведения исследований, направленных на решение |
| отдельных исследовательских задач. Улучшил навыки оформления отчетов |
| и презентаций. |
| ПК-1 Способность решать исследовательские задачи в рамках реализации научного (научно- технического) проекта под руководством более квалифицированного работника.  ПК-2 Способность проводить презентации, научно-технические отчеты и публичную защиту проектных работ, оформлять результаты исследований в виде статей и докладов.  ПК-4 Способность создавать алгоритмы, программные модули и комплексы для решения исследовательских задач в области системного анализа и управления |

1. Характеристика работы обучающегося[[7]](#footnote-7)

|  |
| --- |
| В результате прохождения практики Калинин Д.А. ответственно выполнял |
| все поставленные перед ним задачи. Продемонстрировал умение |
| использования в практической работе теоретических знаний, полученных |
| во время обучения, а также способность к самоорганизации и |
| самообразованию. |
|  |
|  |
|  |
|  |

|  |
| --- |
| 1. Замечания руководителя от Профильной организации о прохождении практики |
| обучающимся |
| Замечаний нет. |
|  |

Руководитель от Профильной организации

директор ИИТК, зав. НИЛ ИИТК

должность подпись расшифровка подписи

**М.П.**

1. Профильной организацией может являться Университет, с указанием подразделения, где проходит практика; [↑](#footnote-ref-1)
2. При проведении практики в Университете обязанности руководителя от профильной организации возлагаются на руководителя структурного подразделения, в котором проводится практика. [↑](#footnote-ref-2)
3. Если практика проводилась в подразделениях Университета, то ставится печать дирекции или структурного подразделения (при наличии); [↑](#footnote-ref-3)
4. Инструктаж обучающихся по ознакомлению с требованиями охраны труда, технике безопасности, пожарной безопасности, а также правилами внутреннего трудового распорядка Профильной организации. В случае практики в подразделении Университета вводный инструктаж проводит руководитель службы охраны труда. [↑](#footnote-ref-4)
5. Инструктаж на рабочем месте проводит ответственный, имеющий соответствующий допуск (удостоверение). При проведении практики в подразделении Университета, инструктаж проводит Руководитель практики с соответствующим допуском. [↑](#footnote-ref-5)
6. Объем и содержание отчета определяется рабочей программой практик и методическими указаниями. [↑](#footnote-ref-6)
7. Оценка работы обучающегося по итогам прохождения практики, информация о квалификации (разряде, категории), присвоенной в период практики с указанием даты присвоения (при наличии), личные и профессиональные качества, проявленные в ходе практики. [↑](#footnote-ref-7)