Учреждение образования
«Белорусский государственный университет культуры и искусств»

Факультет культурологии и социально-культурной деятельности

Кафедра информационных технологий в культуре

СОГЛАСОВАНО
Заведующий кафедрой
_____ Т.С. Жилинская
«05» мая 2025 г.

СОГЛАСОВАНО
Декан факультета
_____ Н.Е. Шелупенко
«30» мая 2025 г.

УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ПО УЧЕБНОЙ ДИСЦИПЛИНЕ
Анализ данных и визуализация в культуре

для специальности 6-05-0314-03 Социально-культурный менеджмент и коммуникации профилизации Мультимедийные технологии и цифровые коммуникации

СОСТАВИТЕЛИ:

Т. И. Песецкая, доцент кафедры информационных технологий в культуре учреждения образования «Белорусский государственный университет культуры и искусств», кандидат физико-математических наук

Рассмотрено и утверждено
на заседании Совета факультета культурологии и социально-культурной деятельности 27 06 2025 г.
протокол № 10

Минск 2025

Учебно-методический комплекс составлен в соответствии с примерным учебным планом по специальности 6-05-0314-03 Социально-культурный менеджмент и коммуникации, утвержденным Первым заместителем Министра образования Республики Беларусь от 30.01.2023 рег. № 6-05-03-013/пр. и учебного плана учреждения высшего образования по специальности 6-05-0314-03 Социально-культурный менеджмент и коммуникации, профилизации «Мультимедийные технологии и цифровые коммуникации» рег. № 6-05-03-26/23уч. от 15.02.2023

СОСТАВИТЕЛЬ:
Т. И. Песецкая, доцент кафедры информационных технологий в культуре учреждения образования «Белорусский государственный университет культуры и искусств», кандидат физико-математических наук

РЕЦЕНЗЕНТЫ:

С.И. Зенько, доцент кафедры информатики и методики преподавания информатики Белорусского государственного педагогического университета им. М.Танка, кандидат педагогических наук, доцент;

В. В. Казаченок, заведующий кафедрой компьютерных технологий и систем факультета прикладной математики и информатики БГУ, доктор педагогических наук, кандидат физико-математических наук, профессор.

РЕКОМЕНДОВАН К РАССМОТРЕНИЮ:
кафедрой информационных технологий в культуре учреждения образования «Белорусский государственный университет культуры и искусств» (протокол № 9 от 22.05.2024)

ОГЛАВЛЕНИЕ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Учебная дисциплина «Анализ данных и визуализация в культуре» предназначена для студентов специальности 6-05-0314-03 Социально-культурный менеджмент и коммуникации, профилизации «Мультимедийные технологии и цифровые коммуникации».

Изучение учебной дисциплины «Анализ данных и визуализация в культуре» направлено на обучение студентов теоретическим основам и методам анализа данных сферы культуры, и практическим подходам визуализации данных и результатов культурологических исследований, необходимым для решения задач сферы культуры, требующих использования инструментария информационных технологий.

Цель изучения дисциплины состоит в обучении использованию методов анализа больших данных (Big Data) и интеллектуального анализа (Data Mining), а также подходов к визуализации данных для решения творческих задач в сфере социально-культурной деятельности.

Для достижения поставленной цели необходимо решение следующих задач:

1) формирование системы базовых знаний в сфере хранения и обработки больших данных, а также основ интеллектуального анализа больших данных;

2) изучение методов и подходов к анализу больших данных, таких как статистический анализ, машинное обучение, кластерный анализ и другие;

3) выработка навыков и умений по использованию программных средств

анализа данных для решения практических задач социально-культурной сферы;

4) изучение специфики и основных этапов интерпретации результатов анализа данных в сфере культуры;

5) освоение подходов и программных средства визуализации больших данных;

6) формирование навыков и умений визуального представления результатов аналитического исследования данных сферы культуры и искусств.

Знания и навыки, полученные при изучении учебной дисциплины «Анализ данных и визуализация в культуре», необходимы при изучении таких учебных дисциплин, как: «Проектирование информационных ресурсов и систем», «Технологии создания баз данных сферы культуры», «Информационные технологии в культуре».

В соответствии с учебным планом учреждения высшего образования по специальности 6-05-0314-03 Социально-культурный менеджмент и коммуникации, профилизации «Мультимедийные технологии и цифровые коммуникации» освоение образовательной программы по учебной

дисциплине «Анализ данных и визуализация в культуре» должно обеспечивать формирование следующей специальной компетенции:

СК-39. Осуществление анализа и визуализации данных в социально-культурной сфере.

В результате изучения учебной дисциплины «Анализ данных и визуализация в культуре» студенты должны *знать*:

– место и роль анализа больших данных и интеллектуального анализа в системе научных знаний;

– методы и подходы к анализу больших данных;

– современные программные средства и технологии анализа больших данных;

– теоретические аспекты интерпретации результатов анализа данных;

– основы визуализации результатов анализа больших данных сферы культуры;

– базовые этапы и правила представления данных в сфере культуры;

Студенты должны *уметь*:

– определять сферу необходимых для анализа данных для принятия решений в сфере культуры;

– использовать методы анализа больших данных для принятия эффективных решений в сфере культуры;

– объективно интерпретировать результаты анализа данных сферы культуры;

– наглядно визуализировать выводы и обоснования для принятия управленческих решений в сфере культуры.

Студенты должны приобрести *навыки*:

– сбора массивов данных из различных источников;

– подбора методов и программных средств анализа больших данных;

– структурирования и сопоставление результатов анализа данных, а также определения причинно-следственных связей и синтеза всей полученной информации для формулирование заключительных выводов;

– визуализации и представления результатов анализа данных в сфере культуры.

Методы и технологии обучения.

На лекциях особое внимание уделяется рассмотрению теоретических аспектов анализа больших данных и примеров практического его использования. Практические занятия направлены на формирование умений практического использования полученных знаний на основе решения конкретных аналитических задач сферы культуры. Семинарские занятия нацелены на разрешение вызовов, которые встречают студенты в процессе решения практических задач и обсуждение практических результатов аналитической деятельности.

Учебным планом на изучение учебной дисциплины «Системный анализ и моделирование информационных процессов» всего предусмотрено

90 часов, из них 44 часа – аудиторные занятия. Примерное распределение аудиторных часов по видам занятий: лекции – 12 часов, практические занятия – 18 часов, семинарские занятия – 14.

Дисциплина рассчитана на один семестр. Текущий контроль осуществляется при выполнении и сдаче отчетов практических работ. Рекомендуемая форма контроля знаний – зачёт.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ

Тема 1. Большие данные и интеллектуальный анализ данных в сфере культуры

- Определение больших данных (Big Data)
- Цифровизация сферы культуры.
- Большие данные сферы культуры
- Интеллектуальный анализ больших данных (Data Mining) и области применения.
- Задачи интеллектуального анализа больших данных
- Этапы интеллектуального анализа
- Интеллектуальный анализ данных в сфере культуры.
- Культурная аналитика и цифровая культура.

Цель: приобрести комплексное представление о больших данных в сфере культуры

В современном мире генерируются огромные объемы данных, и сфера культуры не является исключением. Библиотеки, музеи, архивы, онлайн-платформы – все это источники массивов информации, которые традиционными методами обработать и проанализировать практически невозможно. Большие данные (Big Data) и интеллектуальный анализ данных (Data Mining) предлагают новые возможности для исследования и популяризации культурного наследия, а также для создания новых форм искусства и культурных продуктов.

Представим несколько взглядов на понятие «большие данные» и интеллектуальный анализ данных.

Большие данные – это структурированные и неструктурированные массивы данных большого объема, генерируемые, как правило с высокой скоростью и степенью разнообразия, требующие новых подходов к хранению, обработке и анализу. В сфере культуры это могут быть тексты книг и рукописей, изображения картин и артефактов, аудио- и видеозаписи концертов и спектаклей, данные о посещаемости музеев и библиотек, отзывы пользователей в социальных сетях и многое другое.

Более сложное восприятие больших данных дадим с технической точки зрения. Большие данные – обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях – весь мировой объём данных, и вытекающих из этого трансформационных последствий.

В качестве определяющих характеристик для больших данных традиционно выделяют:

– объём, в смысле величины физического объёма;

– скорость, в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов;

многообразие, в смысле возможности одновременной обработки различных типов структурированных и полуструктурированных данных.

В дальнейшем возникли различные вариации и интерпретации этих признака.

С точки зрения информационных технологий, в совокупность подходов и инструментов изначально включались средства массово-параллельной обработки неопределённо структурированных данных, прежде всего, системами управления базами данных категории NoSQL, алгоритмами MapReduce и реализующими их программными каркасами и библиотеками проекта Hadoop. В дальнейшем к серии технологий больших данных стали относить разнообразные информационно-технологические решения, в той или иной степени обеспечивающие сходные по характеристикам возможности по обработке сверхбольших массивов данных.

Интеллектуальный анализ данных – это процесс извлечения знаний и моделей из больших данных. Он использует различные методы, включая статистические методы, машинное обучение и искусственный интеллект, для выявления скрытых закономерностей, прогнозирования будущих трендов и принятия обоснованных решений.

Интеллектуальный анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных. В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных. Обычно такие закономерности нельзя обнаружить при

традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных.

Эти закономерности и тренды можно собрать вместе и определить как модель интеллектуального анализа данных. Модели интеллектуального анализа данных могут применяться к конкретным сценариям, а именно:

– прогнозирование: оценка продаж, прогнозирование нагрузки сервера или времени простоя сервера;

– риск и вероятность: выбор наиболее подходящих заказчиков для целевой рассылки, определение точки равновесия для рискованных сценариев, назначение вероятностей диагнозам или другим результатам;

– рекомендации: определение продуктов, которые с высокой долей вероятности могут быть проданы вместе, создание рекомендаций;

– поиск последовательностей: анализ выбора заказчиков во время совершения покупок, прогнозирование следующего возможного события;

– группирование: разделение заказчиков или событий на кластеры связанных элементов, анализ и прогнозирование общих черт.

Построение модели интеллектуального анализа данных является частью более масштабного процесса, в который входят все задачи, от формулировки вопросов относительно данных и создания модели для ответов на эти вопросы до развертывания модели в рабочей среде. Этот процесс можно представить как последовательность следующих шести базовых шагов (рис.1):

1) постановка задачи;
2) подготовка данных;
3) изучение данных;
4) построение моделей;
5) исследование и проверка моделей;
6) развертывание и обновление моделей.

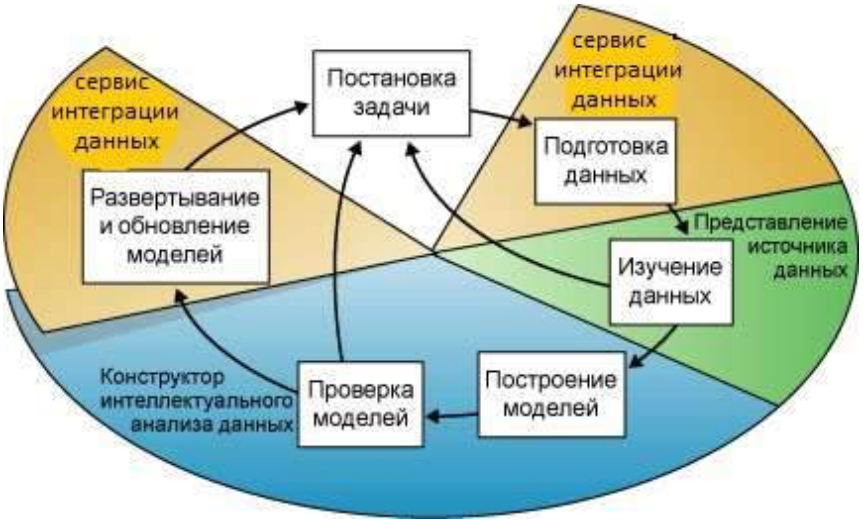

Рисунок 1– Связи между этапами интеллектуального анализа данных

Постановка задачи. Первым шагом процесса интеллектуального анализа данных, является четкое определение проблемы и рассмотрение способов использования данных для решения проблемы.

Этот шаг включает анализ требований, определение области проблемы, метрик, по которым будет выполняться оценка модели, а также определение задач для проекта интеллектуального анализа данных. Эти задачи можно сформулировать в виде следующих вопросов.

Что необходимо найти?

Какие типы связей необходимо найти?

Отражает ли решаемая задача правила или процессы сферы моделирования?

Надо ли делать прогнозы на основании модели интеллектуального анализа данных или просто найти содержательные закономерности и взаимосвязи?

Какой результат или атрибут необходимо спрогнозировать?

Какие виды данных нужно иметь и какого рода информация находится в каждом информационном секторе?

Как связаны информационные сектора?

Нужно ли выполнять очистку, статистическую обработку или обработку, чтобы данные стали применимыми?

Каким образом распределяются данные?

Дают ли данные точное представление о моделируемых процессов?

Также необходимо рассмотреть способы для учета результатов модели в ключевых показателях, например, показателях эффективности.

Вторым шагом процесса интеллектуального анализа данных, как видно из следующей диаграммы, является объединение и очистка данных, определенных во время шага Постановка задачи .

Подготовка данных

Очистка данных – это не только удаление недопустимых данных или интерполяция отсутствующих значений, но и поиск в данных скрытых зависимостей, определение источников самых точных данных и их первичная систематизация, которые больше всего подходят для использования в анализе. Неполные данные, ошибочные данные и входные параметры, которые выглядят как независимые, но на самом деле имеют прочную взаимосвязь, могут непредвиденным образом повлиять на результаты модели.

Поэтому перед началом построения моделей интеллектуального анализа данных следует выявить такие проблемы и определить, как их устранить. Для интеллектуального анализа данных обычно ведётся работа с очень большим набором данных, и невозможно проверить каждую транзакцию на качество данных. Поэтому может потребоваться использовать некоторые виды профилирования данных и автоматизированные средства очистки и фильтрации данных, например, предоставляемые в службах подобных Integration Services (Сервисы интеграции данных) для изучения данных и обнаружения несоответствий.

Изучение данных. Для принятия правильных решений при создании моделей интеллектуального анализа данных необходимо понимать данные. Методы исследования данных включают в себя расчет минимальных и максимальных значений, вычисление средневероятного и стандартного отклонения и изучение распределения данных. Например, по максимальному, минимальному и среднему значениям можно заключить, что выборка данных не является репрезентативной для процессов, и поэтому необходимо получить более сбалансированные данные или изменить предположения, лежащие в основе ожидаемых результатов. Стандартное отклонение и другие характеристики распределения могут сообщить полезные сведения о стабильности и точности результатов. Большая величина стандартного отклонения может свидетельствовать о том, что добавление новых данных поможет усовершенствовать модель. Данные, которые сильно отклоняются от стандартного распределения, могут оказаться искаженными или представлять точную картину реальной проблемы, которая делает сложным подбор соответствующей модели для данных.

Построение моделей. Четвертым шагом процесса интеллектуального анализа данных является построение моделей интеллектуального анализа данных. Знания, полученные при выполнении предыдущих шагов, помогут определить и создать модели.

Сначала определяется структура интеллектуального анализа данных. Структура интеллектуального анализа связана с источником данных, но не содержит никаких данных до обработки. При обработке структуры интеллектуального анализа данных создаются статистические выражения и другая статистическая информацию, которую можно использовать для анализа. Эти данные могут использоваться любой моделью интеллектуального анализа данных, которая основана на этой структуре. Обработку модели часто называют обучением. Обучение обозначает процесс применения некоторого математического алгоритма к данным в структуре с целью выявить закономерности. Закономерности, обнаруженные в процессе обучения, зависят от выбора обучающих данных, выбранного алгоритма и его конфигурации.

После прохождения данных через модель объект модели интеллектуального анализа данных будет содержать сводные данные и закономерности, которые можно запрашивать и использовать для прогнозирования.

Важно помнить, что при любом изменении данных необходимо обновить и структуру, и модель интеллектуального анализа данных. При обновлении структуры интеллектуального анализа данных путем ее повторной обработки извлекаются данные из источника, включая любые новые данные, если источник обновляется динамически, и повторно обновляет структуру интеллектуального анализа данных. Если на этой структуре основаны существующие модели, можно обновить эти модели, что будет означать их повторное обучение с новыми данными, или оставить модели без изменений.

Исследование и проверка моделей. Пятым шагом процесса интеллектуального анализа данных, как видно из диаграммы ниже, является исследование построенных моделей интеллектуального анализа данных и проверка их эффективности.

Перед развертыванием модели в рабочей среде необходимо проверить эффективность работы модели. Кроме того, во время построения модели обычно создается несколько моделей с различной конфигурацией, а затем проверяются все модели, чтобы определить, какая из них обеспечивает лучшие результаты для поставленной задачи и имеющихся данных.

Системы интеллектуального анализа данных предоставляют средства, помогающие разделить данные на наборы данных для обучения и тестирования, чтобы можно было точно оценить производительность всех моделей на основе одних и того же данных. Набор данных для обучения используется в ходе построения модели, а набор проверочных данных – для проверки точности модели путем создания прогнозирующих запросов. Это

секционирование можно выполнить автоматически при построении модели интеллектуального анализа данных. Точность прогнозов, создаваемых моделями, можно проверить при помощи таких средств как диаграмма точности прогнозов и матрица классификации. Чтобы проверить, ограничена применимость модели имеющимися данными или она может использоваться для совершения выводов относительно генеральной совокупности, можно применить статистический метод, называемый перекрестной проверкой, чтобы автоматически создать подмножества данных и проверить модель по каждому подмножеству.

Если ни одна из моделей, созданных при выполнении шага «Построение моделей», не обладает нужной эффективностью, может возникнуть необходимость вернуться к предыдущему шагу процесса и либо изменить постановку задачи либо выполнить повторное изучение данных в исходном наборе данных.

Развертывание и обновление моделей. Последним шагом процесса интеллектуального анализа данных является развертывание наиболее эффективных моделей в рабочей среде.

После развертывания моделей интеллектуального анализа данных в рабочей среде можно выполнять множество задач, соответствующих потребностям пользователя. Ниже перечислены некоторые возможные задачи:

Используйте модели для создания прогнозов, которые можно затем использовать для принятия решений.

Создание запросов содержимого для получения статистики, правил или формул из модели.

Внедрение функций интеллектуального анализа данных непосредственно в цифровые приложения.

Обновление моделей после просмотра и анализа. После любого обновления необходимо выполнить повторную обработку моделей.

Динамическое обновление моделей по мере поступления новых данных и постоянные изменения, направленные на повышение эффективности решения, должны быть частью стратегии развертывания.

Применение больших данных и интеллектуального анализа данных в сфере культуры

Цифровая консервация и каталогизация. С помощью интеллектуального анализа данных можно автоматизировать процессы описания и каталогизации музейных экспонатов, архивных документов и библиотечных фондов, значительно ускоряя и удешевляя эти трудоемкие задачи. Анализ изображений позволяет создавать автоматизированные системы для распознавания объектов и поиска похожих артефактов.

Анализ пользовательского поведения. Анализ данных о посещаемости выставок, предпочтениях читателей, просмотре видео и прослушивании музыки позволяет музеям, библиотекам и другим культурным учреждениям лучше понимать свою аудиторию и адаптировать свою деятельность к ее потребностям.

Персонализация культурных услуг. На основе анализа данных можно создавать персонализированные рекомендации для пользователей, предлагая им культурные мероприятия, книги и фильмы, которые могут им понравиться.

Создание новых форм искусства. Большие данные могут служить источником вдохновения для художников, композиторов и дизайнеров. Анализ данных позволяет создавать интерактивные инсталляции, генеративные произведения искусства и новые формы музыкальных композиций.

Исследование культурных трендов. Анализ данных из социальных сетей, новостных источников и других онлайн-платформ позволяет выявлять и отслеживать культурные тренды, что важно для планирования культурных мероприятий и создания новых культурных продуктов.

Несмотря на огромный потенциал, применение больших данных и интеллектуального анализа данных в сфере культуры сталкивается с рядом вызовов:

– для эффективного анализа необходимы качественные и достоверные данные. В сфере культуры это часто бывает проблематично из-за наличия неполных, неконсистентных и неструктурированных данных;

– анализ данных о пользователях должен осуществляться с учетом требований к защите данных и конфиденциальности;

– применение интеллектуального анализа данных в сфере культуры поднимает ряд этических вопросов, например, вопросы авторского права и ответственности за решения, принятые на основе анализа данных.

Большие данные и интеллектуальный анализ данных открывают перед сферой культуры новые горизонты, позволяя глубже изучать культурное наследие, создавать новые формы искусства и более эффективно взаимодействовать с аудиторией. Однако, необходимо учитывать вызовы и ограничения, связанные с использованием этих технологий, и обеспечивать этическое и ответственное применение технологий анализа больших данных в культурной сфере.

Тема 2. Методы и подходы к анализу больших данных

- Математические методы и компьютерные технологи анализа данных в сфере культуры.
- Статистические методы анализа больших данных.
- Дисперсионный, корреляционный и регрессионный анализ.
- Подходы кластерного анализа.
- Технологии машинного обучения в анализе больших данных.

Цель: рассмотреть подходы к анализу больших данных

К методам и компьютерным технологиям анализа данных в сфере культуры можно отнести общеприменимые методы из других областей применения анализа больших данных, такие как статистические методы, математическое моделирование, глубинный анализ данных, машинное обучение, сетевой анализ и анализ сложных динамических систем.

Для решения задач, связанных с анализом данных (выявление скрытых взаимосвязей внутри массивов данных) при наличии случайных и непредсказуемых воздействий, математиками и другими исследователями за последние двести лет был выработан мощный и гибкий арсенал методов, называемых в совокупности статистическими методами анализа данных. За это время накоплен большой опыт успешного применения этих методов в разных сферах человеческой деятельности, от экономики до космических исследований. И при определенных условиях эти методы позволяют получать оптимальные решения.

Статистические методы (методы, основанные на использовании математической статистики) являются эффективным инструментом сбора и анализа информации. Применение этих методов не требует больших затрат и позволяет с заданной степенью точности и достоверностью судить о состоянии исследуемых явлений (объектов, процессов), прогнозировать и регулировать проблемы на всех этапах их жизненного цикла и на основе этого вырабатывать оптимальные управленческие решения.

К настоящему времени в мировой практике накоплен огромный арсенал статистических методов, многие из которых могут быть достаточно эффективно использованы для решения различных вопросов. Условно все методы можно классифицировать по признаку общности на три основные группы: графические методы, методы анализа статистических совокупностей и экономико-математические методы. Предложенная классификация не является ни универсальной, ни исчерпывающей, но она дает наглядное представление о разнообразии статистических методов и о тех потенциальных возможностях, которыми они располагают по части их использования при анализе данных.

Графические методы основаны на применении графических средств анализа статистических данных. В эту группу могут быть включены такие методы, как контрольный листок, диаграмма Парето, схема Исикавы, гистограмма, диаграмма разброса, расслоение, контрольная карта, график временного ряда и др. Данные методы не требуют сложных вычислений, могут использоваться как самостоятельно, так и в комплексе с другими методами. Овладение ими не представляет особого труда не только для инженерно-технических работников, но и для специалистов низшего звена. Вместе с тем это весьма эффективные методы. Недаром они находят самое широкое применение в промышленности, особенно в работе групп качества.

Методы анализа статистических совокупностей служат для исследования информации, когда изменение анализируемого параметра носит случайный характер. Основными методами, включаемыми в данную группу, являются: регрессивный, дисперсионный и факторный виды анализа, метод сравнения средних, метод сравнения дисперсий и др. Эти методы позволяют установить зависимость изучаемых явлений от случайных факторов как качественную (дисперсионный анализ), так и количественную (корреляционный анализ); исследовать связи между случайными и неслучайными величинами (регрессивный анализ); выявить роль отдельных факторов в изменении анализируемого параметра (факторный анализ) и т.д.

Экономико-математические методы представляют собой сочетание экономических, математических и кибернетических методов. Центральным понятием методов этой группы является оптимизация, т. е. процесс нахождения наилучшего варианта из множества возможных с учетом принятого критерия (критерия оптимальности). Строго говоря, экономико-математические методы не являются чисто статистическими, но они широко используют аппарат математической статистики, что дает основание включить их в рассматриваемую классификацию статистических методов. Для целей, связанных с обеспечением качества, из достаточно обширной группы экономико-математических методов следует выделить в первую очередь следующие: математическое программирование (линейное, нелинейное, динамическое); планирование эксперимента; имитационное моделирование: теория игр; теория массового обслуживания; теория расписаний; функционально-стоимостной анализ и др.

Анализ данных с помощью статистических методов может быть выполнен в несколько этапов (табл. 1).

Таблица 1: Этапы анализа данных и их статистические методы

№ п/п	Этапы анализа данных	Статистические методы исследования
1.	Описание данных	Описательная статистика, определение необходимого объема выборки.
2.	Изучение сходств и различий	Статистические критерии: Крамера-Уэлча, Вилкоксона-Манна-Уитни, хи–квадрат, Фишера и др.
3.	Исследование зависимостей	Корреляционный анализ, дисперсионный анализ, регрессионный анализ.
4.	Снижение размерности	Факторный анализ, метод главных компонент.
5.	Классификация и прогноз	Дискриминантный анализ, кластерный анализ, группировка.

Описание данных. В практических задачах обычно имеется совокупность наблюдений (десятки, сотни, а то и тысячи результатов измерений индивидуальных характеристик), в связи с этим возникает задача компактного описания имеющихся данных.

Для этого используют методы описательной статистики – описания результатов с помощью различных агрегированных показателей графиков. Кроме того, некоторые показатели описательной статистики используются и в других статистических методах.

Для результатов измерений в шкале отношений показатели описательной статистики можно разбить на несколько групп.

Показатели положения – описывают положение экспериментальных данных на числовой оси. Примеры таких данных – максимальный и минимальный элементы выборки, среднее значение, медиана, мода и др.

Показатели разброса – описывают степень разброса данных относительно своего центра (среднего значения). К ним относятся: выборочная дисперсия, разность между минимальным и максимальными элементами (размах, интеграл) выборки и др.

Показатели асимметрии – положение медианы относительно среднего и др.

Гистограмма – показатели используются для наглядного представления и первичного (визуального) анализа результатов измерений характеристик экспериментальной и контрольной групп.

Изучение сходств и различий (сравнение двух выборок) – задача заключается в установлении совпадений или различий характеристик двух имеющихся выборок.

Типовой задачей анализа данных является задача установления совпадений или различий характеристик экспериментальной и контрольной групп. Для этого формулируются статистические гипотезы: гипотеза об отсутствии различий (так называемая нулевая гипотеза) и гипотеза о значимости различий (так называемая альтернативная гипотеза).

Для принятия решения о том, какую из гипотез (нулевую или альтернативную) следует принять, используют решающие правила – статистические критерии. То есть на основании информации о результатах наблюдений (характеристиках членов экспериментальной и контрольной групп) вычисляется число, называемое эмпирическим значением критерия. Это число сравнивается с известным (например, заданным таблично) эталонным числом, называемым критическим значением критерия.

Критические значения приводятся, как правило, для нескольких уровней значимости. Уровнем значимости называется вероятность ошибки, заключающейся в отклонении (не принятии) нулевой гипотезы, когда она верна, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны. Обычно используют уровни значимости 0,05; 0,01; 0,001.

Если полученное исследователем эмпирическое значение критерия а оказывается меньше или равно критическому, то принимается нулевая гипотеза - считается, что на заданном уровне значимости (то есть при том значении критического показателя, для которого рассчитано критическое значение критерия) характеристики экспериментальных и контрольных групп совпадают. В противном случае, если эмпирическое значение критерия оказывается строго больше критического, то нулевая гипотеза отвергается и принимается альтернативная гипотеза – характеристики экспериментальной и контрольной групп считаются различными с достоверностью различий (1-а). Например, если а = 0,05 и принята альтернативная гипотеза, то достоверность различий равна 0,95 или 95%. То есть достоверность различия характеристик – это дополнение до единицы уровня значимости при проверке гипотезы о совпадении характеристик двух независимых выборок. Другими словами, чем меньше эмпирическое значение критерия (чем левее оно находится от критического значения), тем больше степень совпадения характеристик сравниваемых объектов. И наоборот, чем больше эмпирическое значение критерия (чем правее оно находится от критического значения), тем сильнее различаются характеристики сравниваемых объектов.

Исследование зависимостей. Если описательная статистика и статистические критерии позволяют, соответственно, компактно представлять полученные результаты и определять сходства и различия, то следующим этапом анализа данных обычно является исследование зависимостей. Для этих целей применяются корреляционный и дисперсионный анализ (для установления факта наличия или отсутствия зависимости между переменными), а также регрессионный анализ (для нахождения количественной зависимости между переменными).

Корреляционный анализ. Корреляция (Correlation) – связь между двумя или более переменными (в последнем случае корреляция называется множественной). Цель корреляционного анализа – установление наличия или отсутствия этой связи. В случае, когда имеются две переменные, значения которых измерены в шкале отношений, используется коэффициент линейной корреляции Пирсона r, который принимает значения от -1 до +1 (его нулевое значение свидетельствует об отсутствии корреляции). Термин «линейный» свидетельствует о том, что исследуется наличие линейной связи между переменными – если $r(x, y) = 1$, то одна переменная линейно зависит от другой (и наоборот), то есть существуют константы a и b, причем $a > 0$, такие что $y = a x + b$.

Для данных, измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена (он может применяться и для данных, измеренных в интервальной шкале, так как является непараметрическим и улавливает тенденцию – изменения переменных в одном направлении), который обозначается s и определяется сравнением рангов – номеров значений сравниваемых переменных в их упорядочении. Коэффициент корреляции Спирмена является менее чувствительным, чем коэффициент корреляции Пирсона (так как первый в случае измерений в шкале отношений учитывает лишь упорядочение x элементов выборки). В то же время он позволяет выявлять корреляцию между монотонно нелинейно связанными переменными (для которых коэффициент Пирсона может показывать незначительную корреляцию.

Универсальных рецептов установления корреляции между немонотонно и нелинейно связанными переменными на сегодняшний день не существует. Отметим, что большое (близкое к плюс единице или к минус единице) значение коэффициента корреляции говорит о связи переменных, но ничего не говорит о причинно-следственных отношениях между ними.

Дисперсионный анализ. Изучение наличия или отсутствия зависимости между переменными можно проводить и с помощью дисперсионного анализа

Вариационный анализ. Его суть заключается в следующем. Дисперсия характеризует «разброс» значений переменной. Переменные связаны, если для объектов, отличающихся значениями одной переменной, отличаются и значения другой переменной. Значит, нужно для всех объектов, имеющих одно и то же значение одной переменной (называемой независимой переменной), посмотреть, насколько различаются (насколько велика дисперсия) значения другой (или других) переменной, называемой зависимой переменной. Дисперсионный анализ как раз и дает возможность сравнить отношение дисперсии зависимой переменной (межгрупповой дисперсии) с дисперсией внутри групп объектов, характеризуемых одними и теми же значениями независимой переменной (внутригрупповой дисперсией). Другими словами, дисперсионный анализ «работает» следующим образом. Выдвигается гипотеза о наличии зависимости между переменными. Выделяются группы элементов выборки с одинаковыми значениями независимой переменной (число таких групп равно числу попарно различных значений независимой переменной). Если гипотеза о зависимости верна, то значения зависимой переменной внутри каждой группы должны не очень различаться (внутригрупповая дисперсия должна быть мала). Напротив, значения зависимой переменной для различных групп должны различаться сильно (межгрупповая дисперсия должна быть велика). То есть, переменные зависимы, если отношение межгрупповой дисперсии к внутригрупповой (обычно обозначаемое буквой F) велико. Если же гипотеза неверна, то это отношение должно быть мало.

Регрессионный анализ. Если корреляционный и дисперсионный анализ, качественно говоря, дают ответ на вопрос, существует ли взаимосвязь между переменными, то регрессионный анализ предназначен для того, чтобы найти «явный вид» этой зависимости. Цель регрессионного анализа – найти функциональную зависимость между переменными. Для этого предполагается, что зависимая переменная (иногда называемая откликом) определяется известной функцией (иногда говорят – моделью), зависящей от зависимой переменной или переменных (иногда называемых факторами) и некоторого параметра. Требуется найти такие значения этого параметра, чтобы полученная зависимость (модель) наилучшим образом описывала имеющиеся экспериментальные данные. Например, в простой линейной регрессии предполагается, что зависимая переменная y является линейной функцией $y = a\,x + b$ от независимой переменной x. Требуется найти значения параметров a и b, при которых прямая $ax + b$ будет наилучшим образом описывать (аппроксимировать) экспериментальные точки $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Снижение размерности. Часто в результате экспериментальных исследований возникают большие массивы информации. Например, каждый из исследуемых объектов описывается по нескольким критериям (измеряются значения нескольких переменных – признаков). Тогда результатом измерений будет таблица с числом ячеек, равным произведению числа объектов на число признаков. Возникает вопрос, а все ли переменные являются информативными, например, отражают изменения, произошедшие в результате изучаемого воздействия? Исследователю желательно было бы выявить эти существенные переменные (это важно с содержательной точки зрения) и сконцентрировать внимание на них. Кроме того, всегда желательно сокращать объемы обрабатываемой информации (не теряя при этом сути). Статистические методы могут помочь и здесь. Существует целый класс задач статистического анализа – методы снижения размерности – цель которых как раз и заключается в уменьшении числа анализируемых переменных либо посредством выделения существенных переменных, либо построения новых показателей (на основании полученных в результате эксперимента). Но за все (в том числе за агрегирование информации) надо платить – такой платой в задачах снижения размерности является та часть вариации (изменений, дисперсии) исходных показателей, которая объясняется изменениями тех показателей, которые не «остаются» в результате снижения размерности (наименее изменчивые показатели или их комбинации).

Для снижения размерности используется факторный анализ, а основными методами являются кратко рассматриваемый ниже метод главных компонент и многомерное шкалирование.

Метод главных компонент заключается в получении нескольких новых показателей – главных компонент, являющихся линейными комбинациями исходных показателей (напомним, что линейной комбинацией называется взвешенная сумма), полученных в результате эксперимента. Главные компоненты упорядочиваются в порядке убывания той дисперсии, которую они «объясняют». Первая главная компонента объясняет большую часть дисперсии, чем вторая, вторая – большую, чем третья и т.д. Понятно, что чем больше главных компонент будет учитываться, тем большую часть изменений можно будет объяснить.

Преимущество метода главных компонент заключается в том, что зачастую первые несколько главных компонент (одна-две-три) объясняют большую часть (например, 80-90%) изменений большого числа (десятков, а иногда и сотен) параметров. Кроме того, может оказаться, что в первые несколько главных компонент входят не все исходные параметры. Тогда можно сделать вывод о том, какие параметры являются существенными и на них следует обратить внимание в первую очередь.

Классификация. Обширную группу задач анализа данных, основывающихся на применении статистических методов, составляют так называемые задачи классификации. В близких смыслах (в зависимости от предметной области) используются также термины: «группировка», «систематизация», «таксономия», «диагностика», «прогноз», «принятие решений», «распознавание образов».

Выделяются три подобласти теории классификации:

– дискриминация (дискриминантный анализ),

– кластеризация (кластерный анализ)

– группировка.

В дискриминантном анализе классы предполагаются заданными (например, обучающими выборками, для элементов которых известно, каким классам они принадлежат: больной-здоровый, легкая степень заболевания – средняя – тяжелая и т.д.). Задача заключается в том, чтобы вновь появляющийся объект отнести к одному из этих классов. У термина «дискриминация» имеется множество синонимов: диагностика (требуется поставить диагноз из конечного списка возможных диагнозов, если известны определенные характеристики пациента и известно, какие диагнозы ставились пациентам, вошедшим в обучающую выборку), распознавание образов с учителем, автоматическая (или статистическая) классификация с учителем и т.д. Если в дискриминантном анализе классы заданы, то кластеризация и группировка предназначены для выявления и выделения классов. Синонимами являются: построение классификации, таксономия, распознавание образов без учителя, автоматическая классификация без учителя и т.д.

Задача кластерного анализа заключается в выделении по эмпирическим данным резко различающихся групп (кластеров) объектов, которые схожи между собой внутри каждой из групп. При группировке, когда резких границ между кластерами не существует, исследователю приходится самому вводить границы между группами объектов. Задачам классификации (как теоретическим их аспектам, так и опыту успешного решения конкретных прикладных задач посвящена многочисленная литература). Рассматривать их подробно в настоящей работе мы не будем. К задачам прогноза обычно относят и задачи анализа временных рядов.

Анализ временных рядов. Временным рядом называется последовательность чисел – значений некоторого показателя, измеренного в различные моменты времени. Временные ряды используются для описания динамики процессов, например, изменения температуры тела, концентрации определенного вещества в крови и т.д. На основании конечного отрезка временного ряда исследователь должен сделать выводы о свойствах

рассматриваемого процесса и тех механизмах (в рамках статистики – вероятностных механизмах), которые порождают этот ряд. При изучении временных рядов ставятся следующие цели: агрегированное описание характерных особенностей ряда; подбор статистических моделей, описывающих временной ряд; предсказание будущих значений на основании прошлых наблюдений (прогноз динамики); выработка рекомендаций по управлению процессом, порождающим временной ряд. На сегодняшний день существует множество моделей и методов, позволяющих достигать перечисленных выше целей с учетом специфики исследуемого процесса. Эти методы подробно описаны в литературе и реализованы в компьютерных статистических пакетах.

Кластерный анализ. Кластерный анализ представляет собой средство исследования топологической структуры совокупности объектов. Он позволяет разбить множество объектов в признаковом пространстве на классы близких между собой объектов. Обнаруженные этим методом "сгустки объектов", называемые кластерами (классами, таксонами), позволяют сформулировать, в конечном счете, гипотезы о логической структуре совокупности. Кластеры, замечательным образом найденные в первый раз и разумно описанные исследователем, после повторного сбора информации и применения кластерного анализа могут "рассыпаться" из-за случайности выявленной кластерной структуры. Это разрушение может произойти, тогда, когда реальная кластерная структура может отсутствовать вообще (исследуемая совокупность однородна) или когда задано не соответствующее реальности число классов. Среди множества методов кластерного анализа мы рассматриваем лишь методы, позволяющие изучать большие массивы данных, не ограничивающиеся возможностями оперативной памяти ЭВМ.

Машинное обучение. Технология машинного обучения на основе анализа данных берёт начало в 1950 году, когда начали разрабатывать первые программы для игры в шашки. За прошедшие десятилетий общий принцип не изменился. Зато благодаря взрывному росту вычислительных мощностей компьютеров многократно усложнились закономерности и прогнозы, создаваемые ими, и расширился круг проблем и задач, решаемых с использованием машинного обучения.

Чтобы запустить процесс машинного обучение, для начала необходимо загрузить в компьютер Датасет(некоторое количество исходных данных), на которых алгоритм будет учиться обрабатывать запросы. Например, могут быть фотографии собак и котов, на которых уже есть метки, обозначающие к кому они относятся. После процесса обучения, программа уже сама сможет распознавать собак и котов на новых изображениях без содержания меток.

Процесс обучения продолжается и после выданных прогнозов, чем больше данных мы проанализировали программой, тем более точно она распознает нужные изображения.

Благодаря машинному обучению компьютеры учатся распознавать на фотографиях и рисунках не только лица, но и пейзажи, предметы, текст и цифры. Что касается текста, то и здесь не обойтись без машинного обучения: функция проверки грамматики сейчас присутствует в любом текстовом редакторе и даже в телефонах. Причем учитывается не только написание слов, но и контекст, оттенки смысла и другие тонкие лингвистические аспекты. Более того, уже существует программное обеспечение, способное без участия человека писать новостные статьи (на тему экономики и, к примеру, спорта).

Все задачи, решаемые с помощью машинного обучения, относятся к одной из следующих категорий:

- Задача регрессии – прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число (2, 35, 76.454 и др.), к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.
- Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.
- Задача кластеризации – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звёзда, чёрная дыра и т. п.).
- Задача уменьшения размерности – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации (например, сжатие данных).
- Задача выявления аномалий – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие: аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

Основная масса задач, решаемых при помощи методов машинного обучения, относится к двум разным видам: обучение с учителем (supervised

learning) либо без него (unsupervised learning). Однако этим учителем вовсе не обязательно является сам программист, который стоит над компьютером и контролирует каждое действие в программе. «Учитель» в терминах машинного обучения – это само вмешательство человека в процесс обработки информации. В обоих видах обучения машине предоставляются исходные данные, которые ей предстоит проанализировать и найти закономерности. Различие лишь в том, что при обучении с учителем есть ряд гипотез, которые необходимо опровергнуть или подтвердить. Эту разницу легко понять на примерах.

Машинное обучение с учителем. Предположим, в нашем распоряжении оказались сведения о десяти тысячах московских квартир: площадь, этаж, район, наличие или отсутствие парковки у дома, расстояние от метро, цена квартиры и т. п. Нам необходимо создать модель, предсказывающую рыночную стоимость квартиры по её параметрам. Это идеальный пример машинного обучения с учителем: у нас есть исходные данные (количество квартир и их свойства, которые называются признаками) и готовый ответ по каждой из квартир – её стоимость. Программе предстоит решить задачу регрессии.

Ещё пример из практики: подтвердить или опровергнуть наличие рака у пациента, зная все его медицинские показатели. Выяснить, является ли входящее письмо спамом, проанализировав его текст. Это всё задачи на классификацию.

Машинное обучение без учителя. В случае обучения без учителя, когда готовых «правильных ответов» системе не предоставлено, всё обстоит ещё интереснее. Например, у нас есть информация о весе и росте какого-то количества людей, и эти данные нужно распределить по трём группам, для каждой из которых предстоит пошить рубашки подходящих размеров. Это задача кластеризации. В этом случае предстоит разделить все данные на 3 кластера (но, как правило, такого строгого и единственно возможного деления нет).

Если взять другую ситуацию, когда каждый из объектов в выборке обладает сотней различных признаков, то основной трудностью будет графическое отображение такой выборки. Поэтому количество признаков уменьшают до двух или трёх, и становится возможным визуализировать их на плоскости или в 3D. Это – задача уменьшения размерности.

Основные алгоритмы моделей машинного обучения

Дерево принятия решений. Это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом

вероятности наступления того или иного события), эффективность, ресурсозатратность.

Для бизнес-процессов это дерево складывается из минимального числа вопросов, предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

Наивная байесовская классификация. Наивные байесовские классификаторы относятся к семейству простых вероятностных классификаторов и берут начало из теоремы Байеса, которая применительно к данному случаю рассматривает функции как независимые (это называется строгим, или наивным, предположением). На практике используется в следующих областях машинного обучения:

- определение спама, приходящего на электронную почту;
- автоматическая привязка новостных статей к тематическим рубрикам;
- выявление эмоциональной окраски текста;
- распознавание лиц и других паттернов на изображениях.

Метод наименьших квадратов

Всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдёт через точки с нормально распределённым отклонением от истинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.

Логистическая регрессия – это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция (аккумулятивное логистическое распределение). Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных. Это востребовано в следующих ситуациях:

- кредитный скоринг;
- замеры успешности проводимых рекламных кампаний;
- прогноз прибыли с определённого товара;
- оценка вероятности землетрясения в конкретную дату.

Метод опорных векторов. Это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того что объект, находящийся в N-мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью (N – 1), чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так: есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы.

SVM и его модификации помогают решать такие сложные задачи машинного обучения, как сплайсинг ДНК, определение пола человека по фотографии, вывод рекламных баннеров на сайты.

Метод ансамблей. Он базируется на алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе их усреднения или итогов голосования. Изначально метод ансамблей был частным случаем байесовского усреднения, но затем усложнился и оброс дополнительными алгоритмами:

- бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (с математической точки зрения это является улучшающим пересечением);
- бэггинг (bagging) – собирает усложнённые классификаторы, при этом параллельно обучая базовые (улучшающее объединение);
- корректирование ошибок выходного кодирования.

Метод ансамблей – более мощный инструмент по сравнению с отдельно стоящими моделями прогнозирования, поскольку:

- он сводит к минимуму влияние случайностей, усредняя ошибки каждого базового классификатора;
- уменьшает дисперсию, поскольку несколько разных моделей, исходящих из разных гипотез, имеют больше шансов прийти к правильному результату, чем одна отдельно взятая;
- исключает выход за рамки множества: если агрегированная гипотеза оказывается вне множества базовых гипотез, то на этапе формирования комбинированной гипотезы оно расширяется при помощи того или иного способа, и гипотеза уже входит в него.

Алгоритмы кластеризации. Кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере – оказались наиболее схожие между собой элементы.

Кластеризировать объекты можно по разным алгоритмам. Чаще всего используют следующие:

- на основе центра тяжести треугольника;
- на базе подключения;
- сокращения размерности;
- плотности (основанные на пространственной кластеризации);
- вероятностные;
- машинное обучение, в том числе нейронные сети.

Алгоритмы кластеризации используются в биологии (исследование взаимодействия генов в геноме, насчитывающем до нескольких тысяч элементов), социологии (обработка результатов социологических исследований методом Уорда, на выходе дающим кластеры с минимальной дисперсией и примерно одинакового размера) и информационных технологиях.

Метод главных компонент. Метод главных компонент, или PCA, представляет собой статистическую операцию по ортогональному преобразованию, которая имеет своей целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

Практические задачи, в которых применяется PCA, – визуализация и большинство процедур сжатия, упрощения, минимизации данных для того, чтобы облегчить процесс обучения. Однако метод главных компонент не годится для ситуаций, когда исходные данные слабо упорядочены (то есть все компоненты метода характеризуются высокой дисперсией). Так что его применимость определяется тем, насколько хорошо изучена и описана предметная область.

Сингулярное разложение

В линейной алгебре сингулярное разложение, или SVD, определяется как разложение прямоугольной матрицы, состоящей из комплексных или вещественных чисел. Так, матрицу M размерностью $m \times n$ можно разложить таким образом, что $M = U\Sigma V$, где U и V будут унитарными матрицами, а Σ – диагональной.

Одним из частных случаев сингулярного разложения является метод главных компонент. Самые первые технологии компьютерного зрения разрабатывались на основе SVD и PCA и работали следующим образом: вначале лица (или другие паттерны, которые предстояло найти) представляли в виде суммы базисных компонент, затем уменьшали их размерность, после

чего производили их сопоставление с изображениями из выборки. Современные алгоритмы сингулярного разложения в машинном обучении, конечно, значительно сложнее и изощрённее, чем их предшественники, но суть их в целом нем изменилась.

Анализ независимых компонент (ICA)

Это один из статистических методов, который выявляет скрытые факторы, оказывающие влияние на случайные величины, сигналы и пр. ICA формирует порождающую модель для баз многофакторных данных. Переменные в модели содержат некоторые скрытые переменные, причем нет никакой информации о правилах их смешивания. Эти скрытые переменные являются независимыми компонентами выборки и считаются негауссовскими сигналами.

В отличие от анализа главных компонент, который связан с данным методом, анализ независимых компонент более эффективен, особенно в тех случаях, когда классические подходы оказываются бессильны. Он обнаруживает скрытые причины явлений и благодаря этому нашёл широкое применение в самых различных областях – от астрономии и медицины до распознавания речи, автоматического тестирования и анализа динамики финансовых показателей.

Тема 3. Программные средства для анализа данных

- Программные средства обработки статистических данных
- Средства графического анализа
- Извлечение данных инструментами веб-скрейпинга
- Использование искусственного интеллекта для сбора, извлечения и анализа данных

Цель: рассмотреть программные средства для анализа данных

Программные средства для анализа данных можно разделить на несколько категорий:

Статистические пакеты представляют из себя традиционные инструменты, предназначенные для проведения статистического анализа данных. Они предоставляют широкий набор методов для описательной статистики, проверки гипотез, регрессионного анализа и многого другого. К наиболее известным представителям относятся SPSS, SAS, R (с пакетом STATS). Эти пакеты мощны, но часто требуют определенных знаний в статистике.

Инструменты визуализации данных. Визуализация – ключевой этап анализа данных. Графическое представление данных позволяет быстро и эффективно выявлять закономерности и аномалии. Популярные инструменты в этой области: Tableau, Power BI, Qlik Sense. Они отличаются интуитивным интерфейсом и возможностью создавать разнообразные интерактивные графики.

Платформы машинного обучения. К наиболее популярным платформам относятся: Python с библиотеками scikit-learn, TensorFlow, PyTorch, а также облачные сервисы, такие как Amazon SageMaker, Google Cloud AI Platform и Azure Machine Learning. Эти инструменты требуют более глубоких знаний программирования, но обладают огромным потенциалом.

Системы обработки больших данных. Для анализа больших данных (Big Data) используются специализированные платформы, такие как Hadoop, Spark. Они позволяют эффективно обрабатывать и анализировать данные, распределенные на множестве компьютеров.

Специализированные программы. Существуют также программы, ориентированные на анализ данных в конкретных областях, например, программное обеспечение для биоинформатики, анализа текстов, обработки изображений и т.д.

Выбор программного обеспечения зависит от конкретных задач и ресурсов (Таблица 2). При выборе необходимо учитывать:

- Тип данных: структурированные, неструктурированные, временные ряды и т.д.
- Объем данных: небольшие наборы данных, большие данные.
- Необходимый функционал: статистический анализ, машинное обучение, визуализация.
- Уровень владения программированием: некоторые инструменты требуют глубоких знаний программирования, другие – более интуитивны.
- Бюджет: некоторые инструменты бесплатны, другие – коммерческие.

Таблица 2: Сравнительный анализ программного обеспечения

Программное обеспечение	Тип	Преимущества	Недостатки
SPSS	Статистический пакет	Простой интерфейс, широкий функционал	Дорогостоящая лицензия, ограниченные возможности для работы с большими данными
R	Статистический пакет	Бесплатный, открытый код, огромное сообщество, гибкость	Крутой порог вхождения, требует программирования
Python (scikit-learn)	Машинное обучение	Бесплатный, открытый код, гибкость, большое сообщество	Требует программирования
Tableau	Визуализация данных	Простой интерфейс, мощные возможности визуализации	Дорогостоящая лицензия
Power BI	Визуализация данных	Интеграция с Microsoft ecosystem, относительно недорогая	Менее гибкая, чем Tableau

Веб-скрейпинг – это автоматизированный процесс извлечения данных с веб-сайтов. Он позволяет собирать информацию, которая не доступна через API или структурированные форматы данных. Этот метод широко используется в маркетинговых исследованиях, анализе конкурентов, мониторинге цен, сборе новостей и многих других областях. Однако важно помнить о правовых и этических аспектах веб-скрейпинга, включая условия использования сайтов и уважение к роботам.txt.

Существует множество инструментов для ***веб-скрейпинга***, различающихся по сложности, функциональности и методам работы. Рассмотрим некоторые из них:

Библиотеки программирования:

Python: Python – один из самых популярных языков для веб-скрейпинга благодаря своим мощным библиотекам.

Beautiful Soup: Эта библиотека используется для парсинга HTML и XML, позволяя извлекать нужные данные из разметки веб-страницы.

Scrapy: Это высокоуровневый фреймворк для создания веб-скраперов. Он предоставляет инструменты для управления запросами, обработкой данных и сохранения результатов.

Selenium: Эта библиотека позволяет управлять веб-браузером программно, что особенно полезно для работы с динамически генерируемым контентом (JavaScript).

Node.js: Node.js – это платформа для запуска JavaScript на стороне сервера. Для веб-скрейпинга часто используются библиотеки, такие как Cheerio (аналог Beautiful Soup) и Puppeteer (аналог Selenium).

R: Язык R также имеет пакеты для веб-скрейпинга, такие как rvest.

Преимущества использования библиотек программирования: гибкость – можно настраивать скрейпер под ваши конкретные нужды; автоматизация – можно создавать скрипты для автоматического сбора данных; обработка больших объемов данных – библиотеки позволяют эффективно обрабатывать большие объемы данных. Недостатки: требуются навыки программирования, более сложная настройка.

Инструменты с графическим интерфейсом (GUI)

Эти инструменты позволяют создавать веб-скраперы без написания кода. Однако их функциональность обычно ограничена. Примеры таких инструментов:

Octoparse: Популярный инструмент с удобным интерфейсом, позволяющий создавать скрейперы без программирования.

ParseHub: Еще один визуальный инструмент для веб-скрейпинга.

Apify: Платформа, предоставляющая инструменты для создания и запуска веб-скраперов.

Преимущества GUI-инструментов: простота использования (не требует навыков программирования), быстрая настройка – можно быстро создать скрейпер для простых задач. Недостатки: ограниченная функциональность – менее гибкие, чем библиотеки программирования; может быть дорогостоящим, так многие инструменты требуют подписки.

Выбор инструментов зависит от ваших навыков программирования, сложности задачи и доступных ресурсов. Для простых задач можно использовать GUI-инструменты. Для более сложных задач и больших объемов данных обычно требуется использование библиотек программирования. Не забывайте всегда проверять условия использования веб-сайтов и уважать правила роботов.txt перед началом веб-скрейпинга.

Тема 4. Основы и специфика интерпретации результатов анализа анных в сфере культуры

- Понятия интерпретации результатов анализа данных
- Теоретические аспекты интерпретации результатов анализа данных
- Определение личной позиции исследователя
- Этапы интерпретации результатов анализа данных

Цель: изучить основы и специфику интерпретации результатов анализа данных в сфере культуры

Анализ данных – это лишь половина пути к получению знаний. Другая, не менее важная, половина – это интерпретация полученных результатов. Без правильной интерпретации даже самый тщательный анализ данных останется бесполезным набором чисел и графиков. Сегодня мы рассмотрим ключевые аспекты интерпретации, различие между теоретической и эмпирической обработкой данных, а также процесс преобразования результатов статистического анализа в эмпирические и, далее, теоретические знания.

Объяснение и обобщение. Интерпретация результатов анализа данных – это процесс объяснения полученных результатов, выявления скрытых закономерностей и обобщения выводов для более широкого контекста. Она включает в себя:

описание результатов – четкое и лаконичное описание полученных статистических показателей, графиков и таблиц;

объяснение результатов – поиск причинно-следственных связей между переменными, интерпретация статистической значимости результатов;

обобщение результатов – формулировка выводов, которые могут быть применимы к более широкой популяции, чем та, которая использовалась в исследовании;

учет ограничений исследования – выявление ограничений исследования и потенциальных источников погрешностей.

Теоретическая обработка данных и эмпирическая обработка данных. Теоретическая обработка данных – это этап, предшествующий сбору и анализу данных. Он включает в себя формулирование гипотез, определение исследовательских вопросов, выбор методов анализа и разработку теоретической модели. Здесь мы работаем с абстрактными понятиями и концепциями, опираясь на существующие теории и знания. Эмпирическая обработка данных – это работа с реальными данными, собранными в ходе исследования. Она включает в себя сбор, очистку, анализ и интерпретацию данных. Здесь мы работаем с конкретными наблюдениями и измерениями.

Важно отметить, что теоретическая и эмпирическая обработки данных тесно взаимосвязаны. Теоретическая модель направляет эмпирическое исследование, а результаты эмпирического анализа могут подтверждать, опровергать или модифицировать теоретическую модель.

Преобразование данных результата статистического анализа в эмпирические знания. Статистический анализ предоставляет количественные результаты, которые сами по себе не являются знанием. Для преобразования этих результатов в эмпирические знания необходимо:

– контекстуализация – понимание результатов в контексте исследования, учет специфики выборки и метода сбора данных;

– интерпретация статистической значимости – оценка того, насколько полученные результаты являются достоверными и не случайными;

– формулировка эмпирических выводов – создание четких и конкретных выводов на основе результатов анализа, которые описывают наблюдаемые закономерности в данных.

Получение теоретических знаний на базе эмпирических знаний. Эмпирические знания, полученные на основе анализа данных, являются основой для построения и проверки теоретических знаний. Этот процесс включает:

– сравнение с существующими теориями – оценка соответствия полученных эмпирических результатов существующим теоретическим моделям;

– разработка новых теорий – создание новых теоретических моделей для объяснения полученных результатов, если существующие теории не могут их объяснить;

– модификация существующих теорий – внесение изменений в существующие теории на основе полученных эмпирических результатов;

– формулировка обобщений – создание общих закономерностей и принципов на основе эмпирических данных.

Заключение:

Интерпретация результатов анализа данных – это сложный и многогранный процесс, требующий не только знания методов анализа даных, но и глубокого понимания предметной области исследования. Правильная интерпретация позволяет преобразовать результаты анализа в эмпирические и теоретические знания, которые могут быть использованы для принятия обоснованных решений и развития научных знаний. Важно помнить о необходимости критического мышления, объективности и учета ограничений исследования на всех этапах интерпретации.

Теоретические аспекты интерпретации результатов анализа данных

Проверка, уточнение и модификация исходных гипотез. Исходная гипотеза – это предположение, которое мы стремимся проверить с помощью анализа данных. Результаты анализа могут:

– подтвердить гипотезу: если результаты анализа согласуются с гипотезой, это увеличивает степень нашей уверенности в ее правильности. Однако, полное подтверждение гипотезы практически невозможно; всегда остается вероятность ошибки;

– опровергнуть гипотезу: если результаты анализа противоречат гипотезе, это свидетельствует о ее несостоятельности. Нам необходимо пересмотреть исходные предположения и разработать новые гипотезы;

– уточнить гипотезу: результаты анализа могут потребовать уточнения исходной гипотезы. Например, может оказаться, что гипотеза верна лишь при определенных условиях, или необходимо скорректировать формулировку гипотезы, чтобы она лучше отражала полученные данные;

– модифицировать гипотезу: в некоторых случаях исходная гипотеза может оказаться слишком упрощенной или неполной. Результаты анализа могут подсказать, как ее модифицировать, чтобы она лучше соответствовала реальности.

Определение отношений между данными и гипотезами. Важно понимать, что данные не "доказывают" гипотезу. Данные предоставляют эмпирическую информацию, которая может поддерживать или опровергать гипотезу, но не доказывать ее окончательно. Связь между данными и гипотезой описывается с помощью:

– индуктивного вывода – переход от конкретных наблюдений (данных) к общим утверждениям (гипотезам). Этот процесс не является дедуктивным доказательством, а скорее генерализацией;

– вероятностного подхода – оценка вероятности того, что полученные данные были получены при условии истинности гипотезы. Статистические тесты помогают оценить эту вероятность;

– модельных представлений, где гипотезы часто выражаются в виде моделей, которые описывают предполагаемые отношения между переменными. Анализ данных позволяет оценить соответствие модели реальным данным.

Развитие гипотез до уровня теоретических высказываний. Полученные эмпирические выводы должны быть интегрированы в существующую теоретическую базу или использованы для построения новых теорий, что включает:

– объяснение результатов как поиск теоретических объяснений для полученных эмпирических закономерностей;

– формулировку новых гипотез, заключающуюся в разработке новых гипотез на основе полученных результатов;

– интеграцию в существующую теорию, как включение новых результатов в существующие теоретические модели;

– развитие новых теоретических моделей – создание новых теоретических моделей для объяснения новых закономерностей, выявленных в ходе анализа данных.

Объяснение проблем и их решение. Анализ данных часто выявляет не только подтверждения гипотез, но и неожиданные результаты, противоречия и проблемы. Интерпретация результатов включает следующие шаги:

- Идентификация проблем – выявление противоречий, неожиданных результатов и недостатков в данных или методах анализа.
- Поиск причин проблем – понимание причин возникновения выявленных проблем.
- Предложение решений – разработка стратегий для решения выявленных проблем (например, уточнение методики, сбор дополнительных данных, модификация гипотез).
- Пересмотр методологии – В случае серьезных проблем с данными или методами анализа, может потребоваться пересмотр всей методологии исследования.

Личная позиция исследователя оказывает существенное влияние на все этапы исследования – от формулировки проблемы до интерпретации результатов. Сегодня мы поговорим о характеристике исследователя, его компетенциях, умениях и навыках, а также о важности рефлексии и погружения в ситуацию исследования.

Характеристика исследователя: 1) критическое мышление – способность анализировать информацию, выявлять ошибки и оценивать достоверность источников; 2) систематичность – способность организовывать работу, планировать исследования и структурировать данные; 3) настойчивость, так как исследование часто требует длительной и кропотливой работы; 4) объективность – стремление к беспристрастному анализу данных, исключение личных предубеждений (важно отметить, что полная объективность недостижима, но стремление к ней крайне важно); 5) ответственность – осознание ответственности за качество и достоверность результатов исследования; 6) этическая ответственность – соблюдение этических норм, защита прав участников исследования и использование данных в соответствии с этическими принципами.

Компетенции, умения и навыки исследователя: 1) методологические компетенции – знание различных методов исследования, способность выбирать подходящие методы для решения конкретных задач; 2) аналитические компетенции – способность анализировать данные, выявлять закономерности и формулировать выводы; компьютерные компетенции – владение программным обеспечением для анализа данных, обработки информации и создания презентаций.

Рефлексия исследователя. Рефлексия – это осознание и анализ собственных действий, мыслей и чувств в контексте исследования. Она помогает исследователю:

- Осознать свои предрассудки и предубеждения: выявление и нейтрализация возможного влияния личных взглядов на результаты исследования.
- Проанализировать свои методические решения: оценка правильности выбранных методов и инструментов исследования.
- Оценить качество собственной работы: критический анализ всех этапов исследования.
- Учиться на своих ошибках: извлечение уроков из проблем и трудностей, возникших в ходе исследования.

Погружение в ситуацию. Погружение в ситуацию исследования предполагает: глубокое понимание контекста, как тщательный анализ контекста исследования, учет всех важных факторов и особенностей ситуации; эмпатическое понимание в случае, если исследование связано с людьми – важно понимать их чувства, мотивы и опыт; участие в процессах изучаемого объекта для достижения более глубокого понимания.

Этапы интерпретации результатов анализа данных

Структурирование данных. Прежде чем приступать к интерпретации, необходимо привести полученные данные в упорядоченный и понятный вид. Это включает:

- Организацию данных. Систематизация данных в таблицы, графики, диаграммы и другие визуальные формы представления. Выбор наиболее подходящего способа представления зависит от типа данных и поставленных задач.
- Очистку данных. Удаление ошибок, выбросов и пропусков в данных. Важно понимать, какие методы очистки данных были использованы и как они могли повлиять на результаты.
- Преобразование данных. В некоторых случаях может потребоваться преобразование данных для удобства анализа и интерпретации (например, стандартизация, нормализация).

Сопоставление результатов с данными из других источников. Для повышения достоверности интерпретации необходимо сопоставить полученные результаты с данными из других источников. Это может включать:

- Сравнение с результатами других исследований. Проверка согласованности полученных результатов с результатами аналогичных исследований.
- Использование данных из различных баз данных. Интеграция данных из различных источников для получения более полной картины.

- Учет контекстуальной информации. Включение информации о контексте исследования, которая может повлиять на интерпретацию результатов.

Проведение многоаспектного анализа. Интерпретация результатов не должна ограничиваться узким кругом показателей. Необходимо провести многоаспектный анализ, учитывающий различные факторы и точки зрения:

- Анализ различных переменных. Исследование взаимосвязи между различными переменными, выявление корреляций и причинно-следственных связей.
- Учет потенциальных смешивающих факторов. Оценка влияния внешних факторов на полученные результаты.
- Применение различных статистических методов для проверки робастности полученных результатов.

Определение причинно-следственных связей. Один из самых сложных, но и самых важных этапов интерпретации – определение причинно-следственных связей. Важно помнить, что корреляция не равна причинности. Для установления причинно-следственных связей необходимо:

- Изучение механизмов. Понимание механизмов, лежащих в основе выявленных взаимосвязей.
- Учет временной последовательности. Проверка, предшествует ли предполагаемая причина следствию во времени.
- Исключение альтернативных объяснений. Убеждение в том, что выявленная взаимосвязь не обусловлена другими факторами.

Обобщение и синтез всей полученной информации. После проведения всех предыдущих этапов необходимо обобщить и синтезировать всю полученную информацию:

- Формулировка основных выводов. Чёткое и лаконичное описание основных выводов исследования.
- Систематизация информации. Объединение всех результатов в целостную картину.
- Выявление главных закономерностей. Формулировка общих закономерностей и выводов, вытекающих из исследования.

Формулирование заключительных выводов. Заключительные выводы должны быть обоснованы, четкими, лаконичными и соответствовать полученным результатам. Они должны:

- Отвечать на поставленные вопросы. Выводы должны ответить на вопросы, сформулированные в начале исследования.
- Учитывать ограничения исследования. Необходимо отметить ограничения исследования и потенциальные источники ошибок.

- Предлагать рекомендации. Выводы могут содержать рекомендации для дальнейших исследований или практических действий.

Интерпретация результатов анализа данных – это многоэтапный процесс, требующий системного подхода, критического мышления и глубокого понимания предметной области. Правильная интерпретация позволяет превратить сырые данные в ценные знания, способствующие развитию науки и практики.

Тема 5. Основы визуализации и программные средства визуализации данных

- Основные принципы визуализации
- Способы визуализации данных
- Инструменты визуализации данных
- Программные средства визуализации данных

Цель: изучить основы визуализации данных и программные средства визуализации данных

Визуализация данных является важным этапом в процессе изучения и интерпретации данных. Визуализация данных направлена на представление данных в графическом формате, который должен быть интуитивно понятным и лёгким для понимания.

Визуализация данных требует подготовительной работы по очистке и предварительному анализу данных для подготовки «грязных» данных в «чистые данные», используемые для построения графиков и диаграмм. Но даже с подготовленными данными необходимо придерживаться определённых принципов и методологий, чтобы создать полезную, информативную графику.

Основные принципы визуализации данных

Сравнение. Демонстрация сравнения – основа хорошего научного исследования. Доказательства гипотезы всегда связаны с чем-то другим.

Пример утверждения: «Тёмный шоколад улучшает концентрацию внимания и способность к обучению». Важный вопрос в этом утверждении – «по сравнению с чем?» Без сравнения (относительная гипотеза) утверждение бесполезно. Один из способов показать сравнение – контрольная и экспериментальная группы. Люди одной группы будут есть шоколад, люди во второй группе – не будут. Таким образом, вы сможете сравнить влияние

шоколада на концентрацию и способность к обучению на основе результатов теста или путём измерения активности мозга.

При создании графиков для презентации исследования можно составить график для контрольной и экспериментальной групп с помощью «ящика с усами» или гистограммы. Таким образом, читатели получают чёткое представление об эффекте эксперимента.

Причинно-следственная связь и объяснение. Объяснение показывает причинно-следственную связь в размышлениях над исследуемым вопросом.

Возвращаясь к предыдущему примеру, допустим, что испытуемые из экспериментальной группы получили более высокие баллы по тесту, и это показывает, что тёмный шоколад улучшает концентрацию. Важный вопрос: почему это именно так? Этот вопрос важен потому, что он помогает поднять другие вопросы, которые могут либо опровергнуть, либо подкрепить гипотезу на протяжении всего исследования.

Чтобы показать причинно-следственную связь или механизм, можно измерить активность мозга контрольной и экспериментальной групп и построить графики результатов, показав их рядом. С помощью графика тестовых баллов и графика активности мозга можно продемонстрировать причину того, почему принимавшие шоколад испытуемые получили более высокие баллы, т. е. ответ на вопрос, как тёмный шоколад улучшает когнитивные функции.

Данные со многими переменными. Реальный мир сложен, и отношения между двумя событиями обычно нелинейны. Поэтому в исследованиях есть атрибуты или переменные, которые можно измерить. Все эти переменные по-разному взаимодействуют друг с другом. Некоторые из них могут быть путающими, в то время как другие могут быть важными атрибутами, объясняющими взаимосвязь событий.

Так, корреляция не подразумевает причинно-следственной связи. Поэтому не лучшее решение – ограничивать свое исследование только двумя переменными: это приводит к ошибочным выводам. Таким образом, необходимо показать как можно больше данных на своих графиках.

Парадокс Симпсона, парадокс в вероятностной статистике, когда «при объединении групп исчезает тенденция, возникающая в разных группах данных». Чтобы проиллюстрировать:

Две переменные – отрицательная связь (x, y).

Три переменные – положительная связь (x, y, z) (есть путающие переменные).

Инструмент анализа и анализ. Хороший визуализатор данных не ограничивается имеющимися под рукой инструментами для работы с визуализацией. Визуализирующий данные специалист имеет возможность

переключаться от одной формы выражения к использованию нескольких режимов представления. Например, вместо того чтобы создавать отчёты, содержащие только текст, используйте инфографику: изображения, диаграммы, слова, числа и т. д., всё это обогатит информацию

Документирование графиков соответствующими метками, шкалами и источниками данных.

Когда вы впервые смотрите на график, то сначала видите заголовок, а затем метки контекста графика. Без них график не рассказывает ничего. Хорошие отчёты/графики должным образом документируются, при этом каждому графику присваиваются соответствующие шкалы и метки. Источники данных, используемые для создания графиков, также имеют решающее значение.

Содержательность контента. В конечном счёте, независимо от всех вышеперечисленных принципов, без контента, качественного, актуального и целостного, визуализирующая графика будет бесполезна или она будет вводить в заблуждение.

Для качественной визуализации данных важна наглядная подача материала. Ее основные признаки:

Наиболее простая форма. Позволяет четко видеть, сравнивать количественные данные, в том числе в динамике. Ее подтипы: разные виды графиков, диа-, гисто-, спектрограммы, таблицы.

Аналитический тип. Группа форм для разного рода исследований визуализации графических данных. Их особенность – в возможности установления взаимоотношений, тенденций, связей, в т. ч. с помощью геометрики, особых систем координат и др. Примеры: многоосевые гистограммы; полярный и т. п. графики; карты; диаграммы: спагетти, Эйлера и др.

Концептуальность. Представляет связи понятий, идей конкретной предметной области. Сферы использования – IT, мозговой штурм, образовательные программы, проект-менеджмент и т. п. Это концепт-карты, некоторые виды графов, диаграмма Ганта и др.

Стратегический тип визуализации данных. Позволяет анализировать и сравнивать бизнес-модели и их деятельность в целом и по направлениям. Это органиграммы, карты процессов (в т. ч. контрольная карта Шухарта), диаграммы производительности и др.

Метафорический тип визуализации данных. Структурирует данные для лучшей аналитики в образы дерева, пирамиды, иных понятных систем и конструкций: генеалогическое древо; транспортная карта города и т. п.

Комбинированный тип визуализации данных. Формирование сложной композиции нескольких видов визуалов для системного

соотнесения, изучения и выявления связей разнородных данных: та же схема автодвижения с высотами, качеством дорог и др. характеристиками площади.

Таким образом, наглядность соотносится с особенностями сведений и типом их визуализации (данных и информации), целью представления и спецификой аудитории. Всё это накладывается на то, что физиологически зрительное восприятие – базовое у человека, так как:

- почти три четверти сенсорных рецепторов человека расположены в глазах;
- лишь 10 % информирующих чувств не задействуют для этого зрение;
- в работе с визуальными данными заняты почти 50 % нервных клеток головного мозга.
- люди могут воспроизвести 80 % из того, что они восприняли зрением и повторили, 20 % из увиденного и осмысленного (т. е. прочитанного), 10 % – от услышанного;
- когнитивная функция мозга при принятии и анализе наглядно представленных данных используется на 19 % менее активно, чем при восприятии иного формата, однако эффективность деятельности субъекта с визуальной информацией на 17 % выше;
- акцент на подробностях позволяет запоминать лучше на 4,5 %;
- визуализированные данные воспринимаются быстрее текстовых в 60; тыс. раз, поэтому гайд с иллюстрациями исполняется на 323 % эффективнее, чем без них.

Способы и инструменты визуализации данных

Распределение. Частота встречающихся различных значений в наборе данных. Для визуализации подходят такие инструменты как гистограммы, отображающие частоту значений в интервалах, и ящичковые диаграммы (box plots), демонстрирующие квартили, медиану и выбросы. Для дискретных данных можно использовать столбчатые диаграммы. Выбор зависит от типа данных и того, что нужно подчеркнуть.

Сравнение. Соотношение различные групп или категорий данных. Здесь хорошо работают такие инструменты как столбчатые диаграммы (для сравнения количеств), круговые диаграммы (для отображения долей от целого), точечные диаграммы (scatter plots) (для сравнения двух переменных), и группированные столбчатые диаграммы (для сравнения нескольких категорий в разных группах).

Композиция отображает как составляющие части складываются в целое. В качестве инструмента визуализации здесь применяются круговые диаграммы, древовидные диаграммы (treemaps), диаграммы Парето (для отображения распределения по принципу Парето - 20/80), и мозаичные диаграммы.

Непрерывные и дискретные числовые зависимости показывают тенденции изменения одной переменной в зависимости от другой. Для визуализации непрерывных данных используются такие инструменты как точечные диаграммы (scatter plots), которые показывают корреляцию между переменными, и линейные графики, отображающие тренды. Для дискретных данных – точечные диаграммы или столбчатые диаграммы. Линейная регрессия может быть добавлена к точечным диаграммам для отображения тренда.

Временные зависимости отражают изменяется переменной во времени. Здесь используются линейные графики, облачные графики (area charts), гистограммы (для анализа распределения данных по времени) и диаграммы свечей (candlestick charts) (для финансовых данных).

Географические кластеры отражают распределение данных по географической карте. Картограммы (choropleth maps) отображают данные с помощью цвета или штриховки областей, точечные карты (dot maps) показывают концентрацию данных точками, тепловые карты (heatmaps) отображают плотность данных цветом.

Логические структуры показывают как взаимосвязаны различные элементы данных. Здесь используются диаграммы связей (network graphs), древовидные диаграммы, ментальные карты (mind maps) и диаграммы UML (для описания структурных элементов программного обеспечения).

Программные средства визуализации данных:

Google Data Studio. Самый легкий для изучения и использования инструмент, есть интерактивная панель. Сервис может использовать 17 своих коннекторов и больше сотни иных источников для визуализации баз данных партнеров. Это: от Google: Sheets, Реклама, Analytics, Таблицы; от Yandex: Директ и Метрика; Clou SQL, MySQL, PostgreSQL; YouTube Analytics, файлы CSV, Adwords API, Search Console, Attribution 360 и мн. др.

Достоинства. Бесплатен, удобен и понятен даже начинающим. Как понятно по форматам, особо удачно интегрирован в Гугл. Можно индивидуализировать шаблоны под свои требования. Регулярные обновления – так, с этого года внесены: подстройка отчета перед просмотром и выпуском; отслеживание динамики в источниках; визуал можно отправить по почте либо создать имейл-рассылку.

Недостатки. По сравнению с другими сервисами: маленький инструментарий; ограничены способы обработки вычисляемых полей. При превышении предустановленных ограничений некоторые способы взаимодействия с источниками партнерских данных становятся платными.

Power BI. BI-платформа бизнес-аналитики от Microsoft. Один из его сервисов настроен на исследовательскую визуализацию различных данных –

от задач начальника PR-отдела, до потребностей маркетолога и аналитика. Бесплатна, но есть корпоративная Power BI Pro с расширенным функционалом. Ограничений, также как и в первом сервисе, практически нет. Можно использовать материалы из интернета, файлов с различными расширениями, баз данных, систем управления контактами с клиентами (CRM) и др.

Достоинства. Успешно коннектит материалы разных источников. Варианты визуализации собраны в галереи, представляя разные наборы инструментария. Интуитивно понятный интерфейс разработан как для ПК-версии, так и для работы в облаке. Функционал схож с визуализацией данных в электронных таблицах Excel, поэтому удобен для многих.

Сервис интегрируется с разными Microsoft-продуктами, в т. ч. Azure Cloud Service, SQL Server. Эту BI-платформу можно совместить с уже используемыми фирмой приложениями удобного представления данных. Таким образом, этот сервис открывает больше возможностей для визуальной аналитики, чем Google Data Studio.

Недостатки. Часто требуется использование кастомных коннекторов. Могут возникать проблемы в обработке big data и данных из Гугл и Яндекс. Иногда удобнее применять такие бесплатные инструменты выгрузки данных, как Geneport. Некоторым пользователям недостаточно средств обработки и очистки используемых материалов.

Tableau. Может объединять самые разные по формату источников данные, выстраивать очень привлекательную и с широкими возможностями изучения графику. Tableau считают самой крупной и наиболее доступной для любого юзера системой анализа и визуализации данных.

Использует основные источники – такие как разного типа файлы, так и данные БД, облачные материалы – Microsoft Azure и Excel; MySQL и SQL; Google BigQuery; XML и др.

Достоинства. Как и большинство сервисов, работает с данными, легко коннектируя разные источники. Особенность – возможность пользоваться инструментом одновременно группой специалистов в реальном времени. Может отправить/получить результат визуализации по электронному адресу, создать рассылку, опубликовать созданное на сервере.

Удобный даже для новичка, дружелюбный интерфейс можно настроить под свои нужды. Инструмент с дополнительными возможностями совмещения, наложения объектов и богатой галереей. Многие отмечают отличную поддержку и, при желании, общение с большим количеством пользователей.

Недостатки. К минусам сервиса относят необходимость обработки данных перед использованием. Также часто требуется сопровождение либо консультации профильного специалиста.

ChartBlocks. Несложный сервис для онлайн-формирования HTML5-диаграмм, качественно отображаемых как на аппаратах с любыми ОС, так и на разных браузерах. Обрабатывается информация: любой БД, прямых трансляций из разных источников, а также из электронных таблиц.

Достоинства. Этот мастер визуализации числовых данных доступен для использования даже самому неподготовленному юзеру. Есть возможность подстройки под нужды пользователя как цветовых, так и шрифтовых решений. Визуал можно отправить ссылкой или представить на сайте.

Диаграммы адаптированы под любые аппараты с разными ОС и дисплеями, масштабируются, качественно выводятся на распечатку. К плюсам сервиса относят стабильные обновления – так, скоро обещано внедрение живого stream как источника.

Недостатки. Большинство функций предлагается лишь при приобретении пакетов: профессионального или элитного.

Plotly. Платформа, как и большинство из представленных в этом обзоре, имеет бесплатный вариант и платную версию. Позволяет сформировать и подстроить под свои требования множество разных визуалов – от стандартных графиков до оригинальных дашбордов, в том числе по загруженным извне данным. Возможно получение материала из разных источников: Excel-таблиц, баз MySQL, Redshift и др. Платформа может взаимодействовать с созданными на Python, JavaScript, Matlab, R и др. сервисами.

Достоинства. К основным особенностям этого инструмента относят широчайшие возможности индивидуализации практически любой характеристики формируемого визуала – от толщины линии до представления легенды. Платформа в галерее наряду со стандартными 25 активными графиками предлагает поистине уникальные диаграммы. Особенность библиотеки – открытый однострочный код, т. е. возможна доработка шаблонов, причем редактировать за раз можно несколько визуалов. Есть варианты сохранения результатов в виде векторной графики, возможен png-вариант, можно использовать в html на веб-ресурсе.

Недостатки. При необходимости разрешить проблемы, возникающие при работе, обращаются в работающую в Твиттер техническую службу.

Infogram. Достаточно хорошо известная и не требующая особых познаний программа для визуализации данных. Сервис предоставляет услуги разработки интерактивов с разными тарифными планами, среди которых есть

и не требующий оплаты, но с минимумом функциональности. Строить визуалы можно с использованием следующих источников: Excel-таблицы, Гугл- и интернет-карты, изображения gif-формата, базы MySQL, MS SQL Server, PostgreSQL, Amazon Redcliff, Oracle.

Достоинства. Понятный интуитивно даже не имеющим опыта интерфейс. Дополнительно – сопровождающие юзера пошаговые инструкции. Есть возможность донастройки шаблона под свои нужды, а его доработанный вариант можно сохранить в галерее и использовать далее. Визуал с помощью ссылки или кода устанавливается на веб-ресурс, его можно опубликовать в Твиттер или Пинтерест, переслать некоторые форматы в Гугл Драйв или Дропбокс.

Недостатки. Отсутствие возможности работы с кириллицей. Результаты имеют логотип площадки. Сервис предоставляет ограниченные возможности в бесплатной версии.

DataDeck. Возможна синхронизация разного материала и создание дашборда. Средство визуализации данных позволяет проводить веб-анализ, в том числе исследовать показатели конверсии, определять и оценивать динамику времени нахождения пользователя на сайте, сегментировать ЦА, работать с ключевиками и др. Основные показатели даются в реальном времени. Площадка интегрирована с инструментами Google: AdWords, Drive, Analytics, AdSense; с Excel, MailChimp, Slack, а также с Amazon S3, MySQL и MS SQL Server. Пользователям предлагается пробный вариант, помесячные формы оплаты и индивидуально определяемая стоимость приобретения лицензии.

Достоинства. Интерфейс доступен для использования даже самому неподготовленному пользователю. Есть шаблоны. Возможна работа группы пользователей над одним визуалом в реальном времени. Площадка постоянно развивается, предлагая новые способы деятельности.

Недостатки. Достаточно ограниченный набор как коннект-источников, так и визуал-элементов. Нельзя работать с SQL-данными, нет вычисляемых полей.

Тема 6. Представление и визуализация данных в сфере культуры

- Представление данных в сфере культуры: инфографика
- Программные средства и онлайн сервисы создания инфографики
- Программные средства и онлайн сервисы создания презентаций
- Средства и способы демонстрации, размещения и продвижения результатов культурологических исследований в интернет-пространстве.

Цель: рассмотреть подходы к представлению и визуализации данных в сфере культуры.

Одним из эффективных средств представления данных в сфере культуры является инфографика.

Инфографика – это визуальный способ передачи информации. Она представляет собой одну из форм графического дизайна. Большой объем информации оформляется в понятную для восприятия и эстетически приятную иллюстрацию, график, схему и т.д. Инфографика представляет собой самостоятельную единицу информации, в то время как иллюстрации и схемы лишь сопровождают и раскрывают контекст. Инфографика по своему содержанию сопоставима с целым текстом, а схемы, графики и иллюстрации – нет, поскольку вне контекста будет сложно понять, что они демонстрируют.

Как и во многих вопросах, в инфографике есть два принципиально разных подхода: исследовательский и повествовательный.

Исследовательский подход предполагает четкость передачи информации, что характерно для научных работ или бизнеса.

Повествовательный подход характеризуется красочностью и иллюстративностью – это более уместно в журналистике, рекламе, блогах а так же для сферы культуры.

Функционально инфографику можно разделить на следующие виды:

Статистика. Этот вид инфографики помогает преобразовать сухие факты, сложную статистику и море цифр в красочные и ёмкие картинки, понятные большинству. Но важно сохранить баланс и не перегрузить изображение цифрами, иначе это может оттолкнуть потенциального читателя. Поэтому чем проще, тем лучше. Статистическая инфографика может быть использована как в серьезных, так и в развлекательных целях.

Процесс. Если вам нужно сделать инструкцию или алгоритм действий более понятным и наглядным, то нет лучшего решения, чем изобразить это в инфографике. Вы можете оформить текст в картинки и символы, сопроводить его указателями, стрелками, или пронумеровать – тогда получите схематичное и нативно понятное изображение того, как нужно действовать в той или иной ситуации.

География. С этим видом инфографики наверняка сталкивался каждый, потому что карты в телефоне – это тоже инфографика. Этот вид помогает визуализировать местоположение чего угодно благодаря карте, линиям, геометрическим фигурам и цветовым акцентам, штрихам и т.д.

Таймлайн или шкала времени в инфографике помогают продемонстрировать хронологию каких-либо событий. Обычно изображается прямая или изогнутая линия, направленная вертикально или

горизонтально, а рядом с ней помещаются разъясняющие элементы – символы, текст, картинки. При правильной расстановке акцентов такая инфографика отлично фокусирует внимание читателя на важных событиях.

Сравнение Сравнивая какие-либо объекты, вы можете наглядно их изобразить и сопоставить факты о каждом – для этого, как правило, изображаются колонки, а в них указываются сами объекты сравнения и их свойства. То, на что стоит обратить внимание, можно выделить другим цветом.

Иерархия. Иерархическая инфографика помогает продемонстрировать устройство и порядок какой-либо системы, структуры предприятия, взаимосвязь элементов этой структуры. Также с помощью иерархической инфографики можно выделить приоритетные элементы – наиболее важное поместить в центр или наверх пирамиды, наименее важное – по краям или внизу.

Конспект В инфографике можно зафиксировать важные вещи для лучшего запоминания. Например, рецепт приготовления какого-либо блюда, список покупок в магазине, информацию по обучению и так далее. Также можно составить список любимых фильмов, книг или полезных советов. Список необязательно должен быть последовательным.

Преимущества инфографики:

Заметность – картинка с инфографикой гораздо сильнее привлекает внимание человека, в отличие от массивов текста. Из этого следует, что вашу информацию заметят скорее, если она уместится в одну картинку.

Удобство – если человеку комфортно воспринимать информацию, которой вы с ним делитесь, то он скорее всего к вам вернется. Качественная инфографика, как уже было отмечено, делает информацию очень удобной для восприятия.

Скорость распространения – картинками очень удобно делиться в соцсетях, они быстро распространяются по интернету.

Уникальность – вы можете разработать неповторимый дизайн для своей картинки, чем также запомнитесь читателям и выделитесь среди других.

***Недостатки инфографики*:**

Скорость создания – инфографика является результатом кропотливой работы дизайнера.

Упрощение – чтобы уместить большой массив текста в картинку нужно сильно сократить его, поэтому есть шанс упустить важную информацию.

Чтобы создать качественную инфографику необходимо следовать следующим правилам:

Не перегружать инфографику текстом, так как основную информацию несёт изображение. Чтобы проверить, выполняет ли инфографика свою роль, можно убрать из неё весь текст. Даже без текста должно быть примерно понятно, о чём речь.

Не размещать много текста сплошной «простынёй» – это сильно затрудняет восприятие материала. При этом не необходимо добавлять фразы, чтобы объяснить детали.

Исключать элементы, не несущие смысловой нагрузки – они затрудняют восприятие. Инфографика сильна именно концентрированностью подаваемой информации. Каждая линия, стрелка или значок должны нести смысл. Из-за большого количества элементов, не несущих смысловой нагрузки, эту инфографику сложно воспринимать

Не перегружайть. Важно, чтобы человек легко воспринимал приведённую информацию и не запутался в чрезмерном количестве блоков, картинок и стрелочек.

Выстраивать чёткую и логичную структуру. Наиболее важные компоненты или крупные группы размещаются, как правило, по центру.

Уделять внимание правильному использованию цвета. Следите, чтобы текст не терялся на фоне картинок.

Использовать общедоступную символику и картинки. Они должны быть интуитивно понятны аудитории, вызывать у неё стойкие ассоциации. Например, красный цвет у большинства ассоциируется с запретом, воспринимается как призыв к осторожности или знак опасности. В инфографике с помощью этого цвета можно показать, что чего-то недостаточно (значение ниже нормы).

Оформить все элементы инфографики в одном стиле.

В Таблице 4 приведем онлайн-сервисы для создания инфографики.

Таблица 4: Онлайн сервисы для создания инфографики

Сервис	Описание	Сложность использования	Стоимость	Сильные стороны	Слабые стороны

Canva	Универсальный сервис для дизайна, включая инфографику. Множество шаблонов, простой интерфейс.	Легкая	Бесплатный/ Платные	Огромная библиотека шаблонов, простой интерфейс, многофункциональность, русскоязычный	Некоторые функции/элементы платные
Venngage	Специализируется на профессиональных инфографиках и отчетах. Много шаблонов, инструменты для работы с данными.	Средняя	Бесплатный/ Платные	Профессиональный дизайн, возможности для работы с данными	Менее интуитивный интерфейс, чем Canva
Piktochart	Широкий выбор шаблонов, инструменты для работы с графикой и данными, совместная работа.	Средняя	Бесплатный/ Платные	Множество шаблонов, удобство работы с данными, возможность совместной работы	Функционал может быть ограничен в бесплатной версии
Visme	Создание интерактивных инфографик, презентаций и видео. Мощные инструменты, анимация.	Сложная	Бесплатный/ Платные	Интерактивность, расширенные возможности анимации и визуализации, мощный функционал	Сложный интерфейс, требуется время на освоение
Easel.ly	Простой сервис для быстрого создания инфографики. Много шаблонов, простой интерфейс.	Легкая	Бесплатный/ Платные	Простота использования, скорость создания инфографики	Ограниченный функционал

Infogram	Создание инфографик, диаграмм, отчетов. Быстрый и удобный интерфейс.	Легкая - Средняя	Бесплатный/ Платные	Быстрое создание инфографик, удобный интерфейс	Функционал может быть ограничен в бесплатной версии
Creately	Специализация на диаграммах и визуализации данных. Много типов диаграмм.	Средняя - Сложная	Бесплатный/ Платные	Богатый выбор диаграмм, точная визуализация данных	Менее удобен для создания сложных композиций, не столько ориентирован на инфографику

Онлайн-сервис для создания презентаций – это программа для визуализации данных, создания графических презентаций и видео презентаций доступ к которой осуществляется в сети интернет. Ниже в Таблице 3 представлены наиболее эффективные сервисы для создания презентаций в сфере культуры.

Таблица 3: Онлайн сервисы для создания презентаций

Программное обеспечение	Краткое описание	Платформы	Дополнитель-ные возможности
Microsoft PowerPoint	Простое в использовании, множество функций, доступно на нескольких платформах.	Мобильные устройства, Интернет, ПК	
Keynote	Интуитивно понятный интерфейс, создание привлекательных презентаций, интеграция с PowerPoint.	Mac, iPhone	Интеграция с PowerPoint
Google Презентации	Бесплатный доступ, гибкость, совместная работа, шаблоны.	Интернет	Совместная работа, шаблоны, встроенное видео и анимация
Prezi	Креативная альтернатива PowerPoint,	Веб, ПК	Интерактивные и видео

	интерактивные и видео презентации.		презентации
Piktochart	Онлайн-конструктор, широкий выбор шаблонов, пробная версия.	Интернет	Шаблоны, дополнения, пробная версия
Emaze	Широкий спектр инструментов для дизайна, создание простых слайдов и видео презентаций.	Веб, Мобильные устройства	Пробная версия, мобильная версия
Genially	Универсальный онлайн-инструмент, интерактивные элементы, анимация, шаблоны.	Интернет	Интерактивные элементы, анимация, шаблоны (более 1100 в платной версии)
Beautiful.ai	ИИ для дизайна, умные шаблоны слайдов.	Интернет	ИИ для дизайна, шаблоны
Canva	Популярный онлайн-сервис, огромная библиотека шаблонов, простой интерфейс, бесплатная и платная версии. Подходит для создания презентаций, инфографики, постов в соцсетях и др.	Интернет, Мобильные приложения	Огромная библиотека шаблонов, простой интерфейс, бесплатная и платная версии

Выбор платформы представления результатов культурологических исследований зависит от характера исследования, целевой аудитории и целей продвижения. Рассмотрим несколько вариантов:

Университетские репозитории и электронные журналы: Традиционный и авторитетный способ публикации научных работ. Обеспечивает доступ к результатам для академического сообщества, но может иметь ограниченную аудиторию.

Открытые научные платформы (Open Access) предоставляют свободный доступ к публикациям, повышая видимость исследований. Примеры: arXiv, Zenodo. Важно учитывать репутацию платформы и критерии отбора публикаций.

Персональные веб-сайты и блоги позволяют представлять результаты в более доступной и гибкой форме, используя различные визуальные средства и форматы.

Социальные сети Эффективный инструмент для распространения кратких резюме исследований, визуализации ключевых выводов и взаимодействия с аудиторией. Выбор платформы зависит от целевой аудитории (Twitter, Facebook, Instagram, LinkedIn).

Интерактивные онлайн-платформы позволяют представлять интерактивные визуализации данных, карты, графики, что усиливает вовлеченность аудитории. Примеры: Tableau Public, Flourish.

МАТЕРИАЛЫ ДЛЯ СЕМИНАРОВ

Тема 1. Большие данные и интеллектуальный анализ данных в сфере культуры

Семинар 1. Цифровые технологии и большие данные в сфере культуры 2 часа

Цель: рассмотреть влияние цифровых технологий на современную культуру и проанализировать подходы исследования культурных феноменов в контексте больших данных.

Тематические вопросы к семинару:

- Цифровая культура.
- Цифровизация сферы культуры: достижения, вызовы, перспективы.
- Большие данные сферы культуры, их особенности и специфика хранения.
- Интеллектуальный анализ данных в сфере культуры.
- Культурная аналитика: цифровые методы анализа данных.
- Подходы к изучению культурных феноменов и процессов цифровой среды.

Тема 2. Методы и подходы к анализу больших данных

Семинар 1. Математические методы и компьютерные технологи анализа данных в сфере культуры
2 часа

Цель: рассмотреть и проанализировать математические методы и компьютерные технологи анализа данных в сфере культуры

Тематические вопросы к семинару:

- Математические методы анализа данных в сфере культуры: статистические методы.
- Математические методы анализа данных в сфере культуры: математическое моделирование.
- Математические методы анализа данных в сфере культуры: кластерный анализ.
- Математические методы и компьютерные технологи анализа данных в сфере культуры: сетевой анализ.
- Математические методы и компьютерные технологи анализа данных в сфере культуры: анализ сложных динамических систем.

- Компьютерные технологи анализа данных в сфере культуры: машинное обучение

Тема 3. Программные средства для анализа данных

Семинар 1. Технологии добычи и извлечения данных в сфере культуры
2 часа

Цель: рассмотреть технологии добычи и извлечения данных в сфере культуры.

Тематические вопросы к семинару

- Извлечение данных: инструменты веб-скрейпинга.
- Извлечение данных: извлечение данных документов.
- Извлечение данных: извлечение информации из неструктурированных текстовых источников.
- Методы извлечения данных: парсинг PDF-файлов.
- Методы извлечения данных: синтаксический анализ документов.
- Методы извлечения данных: обработка естественного языка (NLP).
- Использование искусственного интеллекта для сбора, извлечения и анализа данных.

Тема 4. Основы интерпретации результатов анализа данных и специфика в сфере культуры

Семинар 1. Теоретические аспекты интерпретации результатов анализа данных
2 часа

Цель: рассмотреть теоретические аспекты интерпретации результатов анализа данных.

Тематические вопросы к семинару

- Теоретические аспекты: проверка и уточнение гипотез.
- Теоретические аспекты: модификация исходных гипотез.
- Теоретические аспекты: определение отношений между данными и гипотезами.
- Теоретические аспекты: развитие гипотез до уровня теоретических высказываний.
- Специфика интерпретации данных в сфере культуры: проблема перевода качественного контекста в числовой эквивалент.

• Специфика интерпретации данных в сфере культуры: влияние позиции наблюдателя исследователя на интерпретацию числовых данных в формулировании выводов.

Тема 5. Основы визуализации и программные средства визуализации данных

Семинар 1. Основы визуализации данных сферы культуры 2 часа

Цель: изучить основы визуализации данных сферы культуры

Тематические вопросы к семинару

- Особенности визуализации данных сферы культуры
- Моделирование связей между информационными блоками.
- Создание визуальных образов неструктурированной информации.
- Визуализация данных с эффектами анимации.
- Использование интерактивных инструментов для представления данных сферы культуры.
- Разработка и дизайн инфографики для визуализации данных сферы культуры.

Тема 6. Представление и визуализация данных в сфере культуры

Семинар 1. Принципы представления данных в сфере культуры 2 часа

Цель: изучить основные принципы представления данных в сфере культуры

Тематические вопросы к семинару

- Принципы представления данных в сфере культуры: эффективность восприятия.
- Принципы представления данных в сфере культуры: очевидность новизны материала.
- Принципы представления данных в сфере культуры: иерархия уровней понимания информации.
- Принципы представления данных в сфере культуры: простота передачи сложных идей и закономерностей.
- Принципы представления данных в сфере культуры: информативность.

- Принципы представления данных в сфере культуры: кратчайший путь демонстрации выводов.
- Принципы представления данных в сфере культуры: эстетика.

Семинар 2. Целеполагание визуализации и представления данных в сфере культуры
2 часа

Цель: изучить основные подходы для эффективной визуализации и представления данных культурологических исследований

Тематические вопросы к семинару

- Целеполагание визуализации и представления данных в сфере культуры: Наглядность представления данных культурологического контекста.
- Объяснимость идеи и содействие пониманию культурологического исследования.
- Демонстрация организованности и связности данных культурологического контекста.
- Доступность представления выводов культурологического исследования.
- Привлечение внимания и побуждение интереса к культурологическому исследованию.

МАТЕРИАЛЫ ДЛЯ ПРАКТИЧЕСКИХ РАБОТ И УПРАВЛЯЕМОЙ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Тема 1. Большие данные и интеллектуальный анализ данных в сфере культуры

Практическая работа 1. Скрытые связи и закономерности данных 2 часа

Цель: научиться определять скрытые связи и закономерности данных

Задания:

1. Определить тему исследования, взяв за основу направление «Влияние цифровых технологий на культуру современного общества» и уточнив его конкретной цифоровой технологией, например, «Социальные сети», тогда итоговая формулировка темы будет звучать следующим образом «Влияние цифровых технологий на культуру современного общества: социальные сети»

2. Выявить 5 скрытых связей и закономерности вилияния данной цифровой технологии на культуру, например

Формирование виртуальной идентичности и её влияние на реальную жизнь. Социальные сети позволяют создавать и поддерживать виртуальные персоны, которые могут существенно отличаться от реальной идентичности. *Скрытая связь* заключается в том, что эта виртуальная идентичность становится всё более значимой и всё сильнее влияет на формирование реальной личности, создавая иногда искаженное представление о себе и окружающих.

Алгортимная фильтрация и создание "эхо-камер". Алгоритмы социальных сетей формируют "эхо-камеры", предлагая пользователям контент, подтверждающий уже существующие убеждения. *Скрытая связь* – невидимое манипулирование восприятием реальности через персонализированные потоки информации, формирующее однобокое мировоззрение.

3. Выявить кластеры и сегменты общества в которых это влияние действует по разному, например, возрастные кластеры для «Социальных сетей»

4. Составить опросник для исследования культурного феномена по теме исследования и создать его в Google Form.

Тема 2. Методы и подходы к анализу больших данных

Самостоятельная работа 1. Сбор данных для исследования феноменов сферы культуры
8 часов

Цель: Осуществть сбор данных для исследования фенома сферы культуры

Задание:

1. Собрать данные опроса с помощью Google Form по теме исследования.
2. Собрать другие необходимые данные, в базах данных открытого доступа.
3. Описать шкалы измерений, генеральную совокупность и выборку.
4. Подготовить базу данных для их предварительной обработки.

Практическая работа 1. Статистические методы анализа данных: Описательная статистика
2 часа

Цель: научиться применять методы описательной статистики для анализа данных

Задание:

1. С помощью методов описательной статичтики провести предварительный анализ данных исследования на выбранную тему.
2. Сформулировать рабочую гипотезу.

Практическая работа 2. Статистические методы анализа данных: корреляционный и регрессионный анализ
2 часа

Цель: научиться применять методы корреляционного анализа данных

Задание:

1. Выделить кластеры данных.
2. Провести корреляционный и регрессионный анализ данных.

Самостоятельная работа 2. Статистическая гипотеза и статистический вывод
4 часа

Цель: научиться формулировать статистические гипотезы и выводы

С учетом результатов практических работ 1 и 2 оформить отчет, собержащий базу данных с описанием и результаты кореляционного и регресионного анализа. Отчет должен содержать описание шкал измерений, генеральной совокупности и выборки; статистическую гипотезу и статистический вывод, уровень достоверности.

Тема 3. Программные средства для анализа данных

Практическая работа 1. Онлайн сервисы анализа данных
2 часа

Цель: научиться использовать онлайн сервисы для анализа данных

Задание:

С помощью онлайн сервисов анализа данных построить необходимые сводные таблицы, графики, диаграммы, блок-схемы, иллюстрирующие результаты обработки данных.

Самостоятельная работа 1. Концептуальные карты и картографические изображения для анализа данных.
4 часа

Цель: научиться использовать концептуальные карты и картографические изображения для анализа данных

Задание:

Создать концептуальные карты и картографические изображения, иллюстрирующие результаты исследования.

Практическая работа 2. Использование нейронных сетей для анализа данных
2 часа

Цель: научиться использовать нейронные сети для анализа данных

Задание:

1. С помощью нейросетей пополнить исследование дополнительными данными.
2. Используя нейросети, выявить как новые данные дополняют, уточняют, рабочую гипотезу.

Самостоятельная работа 2. Уточнение гипотез
8 часов

Цель: научиться уточнять гипотезы и дополнять новыми результатами анализа данных исследование.

Задание:

3. Провести дополнительные статистические исследования с новыми данными, полученными с помощью нейросетей, и уточненными гипотезами в практической работе 2.

4. Оформить отчет, содержащий графическое представление анализа данных (сводные таблицы, графики, диаграммы, блок-схемы, карты и картографические изображения), а также расширенный набор данных, уточненные гипотезы и результаты уточняющего анализа данных.

Тема 4. Основы интерпретации результатов анализа данных и специфика в сфере культуры

Практическая работа 1. Этапы интерпретации результатов анализа данных
2 часа

Цель: с помощью средств информационных технологий осуществить интерпретацию данных исследования

Задание:

С использованием поисковых систем и нейросетей:

1. Осуществить структурирование данных – результатов исследования.

2. Определить определение причинно-следственных связей. сформулировать заключительные выводы.

Самостоятельная работа1. Обобщение и синтез результатов исследования
5 часов

Цель: с помощью средств информационных технологий осуществить обобщение и синтез результатов исследования

I. С использованием поисковых систем и нейросетей:

1. Сопоставить результаты исследования с данными из других источников.

2. Осуществить обобщение и синтез всей информации, полученной в ходе исследования.

II. Создать презентацию, дающую представление об обобщении результатов иследования.

Тема 5. Основы визуализации и программные средства визуализации данных

Самостоятельная работа 1. Инфографика как инструмент визуализации данных
6 часов

Цель: научиться визуализировать данные сферы культуры с помощью инфографики

Задание:

Разработать дизайн инфографики для визуализации данных исследования, использовав представление данных в виде отношений, распределений, сравнений, отобразив непрерывные и дискретные числовые зависимости, временные зависимости, географические кластеры, логические структуры.

Практическая работа 1. Онлайн сервисы визуализации данных для создания инфографики
2 часа

Цель: научиться применять онлайн сервисы визуализации данных для создания инфографики

Задание:

С помощью онлайн сервисов визуализации данных создать инфографику для визуализации данных исследования, использовав представление данных в виде отношений, распределений, сравнений, отобразив непрерывные и дискретные числовые зависимости, временные зависимости, географические кластеры, логические структуры.

Самостоятельная работа 2. Анимация и интерактивные возможности в представлении данных сферы культуры
6 часов

Цель: научиться применять анимацию и интерактивные средства в представлении данных сферы культуры

Задание:

1. Разработать интерактивный инструментарий для презентации результатов исследования.

2. Создать короткий анимационный ролик, поясняющий результаты исследования.

3. Подготовить отчет, содержащий пояснения по дизайну инфографики, анимационному ролику и интерактивному контенту для представления и визуализации данных исследования.

Практическая работа 2. Онлайн сервисы для создания презентаций данных результатов исследования
2 часа

Цель: научиться применять онлайн-сервисы для создания презентаций данных результатов исследования

Задание:

Создать презентацию, визуализирующую данные исследования, включив в нее элементы анимационный ролик, поясняющий результаты исследования и интерактивные опции для взаимодействия с аудиторией.

Тема 6. **Представление и визуализация данных в сфере культуры**

Практическая работа 1. Презентация данных исследования в сфере культуры
2 часа

Цель: научиться делать презентацию данных исследования в сфере культуры

Задание:

1. В иалых группах представить презентаци. данных исследования в сфере культуры, использовав призентацию, разработанную в предыдущей теме.

2. Проанализировать призентации и доклад выступающего, согласно требованиям предявляемым к визуализации данных в сфере културы, учитывая дизайн ифографики, анимацию и интерактив

Самостоятельная работа 1. Отчет как форма представления данных исследования
5 часов

Цель: научиться оформлять результаты исследования в сфере культуры в виде отчета

Задание:

Составить итоговый отчет по теме исследования, включив в него инфографику и ответы на критические замечания коллег по теме исследования.

СОДЕРЖАНИЕ КУРСА

Тема 1. Большие данные и интеллектуальный анализ данных в сфере культуры

Определение больших данных (Big Data). Современные подходы к обработке и хранению больших данных. Цифровизация сферы культуры. Большие данные сферы культуры. Особенности Больших данных сферы культуры.

Интеллектуальный анализ больших данных (Data Mining) и области применения. Задачи интеллектуального анализа больших данных: прогнозирование и предсказание; классификация и сегментация; выявление скрытых связей и закономерностей. Этапы интеллектуального анализа: предварительный анализ данный; применение алгоритмов машинной обработки данных; интерпретация результатов; применение результатов анализа для принятия решений.

Интеллектуальный анализ данных в сфере культуры. Культурная аналитика и цифровая культура. Подходы к изучению культурных феноменов и процессов цифровой среды.

Тема 2. Методы и подходы к анализу больших данных

Математические методы и компьютерные технологи анализа данных в сфере культуры: статистические методы, математическое моделирование, глубинный анализ данных, машинное обучение, сетевой анализ и анализ сложных динамических систем.

Основные понятия статистики: шкалы измерений; генеральная совокупность и выборка; случайная величина и ее распределение. Статистическая гипотеза. Статистический вывод. Уровень достоверности. Описательная статистика. Статистические критерии. Дисперсионный, корреляционный и регрессионный анализ.

Машинное обучение. Алгоритмы, обучающиеся по данным. Машинное обучение с «учителем» и «без учителя». Метрический и линейный классификаторы. Поиск ассоциативных правил и свойство анти-монотонности. Оценка качества обучения модели.

Кластерный анализ. Основные цели кластеризации. Типы кластерных структур. Методы кластеризации: метод К-средних, метод С-средних, сеть Кохонена.

Тема 3. Программные средства для анализа данных

Программные средства обработки статистических данных: табличные процессоры; специальные статистические компьютерные программы; язык программирования для статистической обработки данных; онлайн сервисы.

Средства графического анализа: построители диаграмм; конструкторы блок-схем и концептуальных карт; географические и картографические инструменты.

Извлечение данных: инструменты веб-скрейпинга; инструменты извлечения базы данных; извлечение данных документов; извлечение информации из неструктурированных текстовых источников. Методы извлечения данных: парсинг PDF-файлов, запросы к базе данных, синтаксический анализ документов, оптическое распознавание символов (OCR), обработка естественного языка (NLP).

Использование искусственного интеллекта для сбора, извлечения и анализа данных.

Тема 4. Основы интерпретации результатов анализа данных и специфика в сфере культуры

Понятия интерпретации результатов анализа данных: объяснение и обобщение; теоретическая обработка данных vs эмпирическая обработка данных; преобразование данных результата статистического анализа в эмпирические знания; получение теоретических знаний на базе эмпирических знаний.

Теоретические аспекты интерпретации результатов анализа данных: проверка, уточнение, модификация исходных гипотез; определение отношений между данными и гипотезами; развитие гипотез до уровня теоретических высказываний; объяснение проблем и их решение.

Определение личной позиции исследователя: характеристика исследователя; описание компетенций, умений и навыков исследователя; рефлексия исследователя и погружения в ситуацию.

Этапы интерпретации результатов анализа данных: структурирование данных; сопоставление результатов с данными из других источников; проведение многоаспектного анализа; определение причинно-следственных связей; обобщение и синтез всей полученной информации; формулирование заключительных выводов.

Специфика интерпретации данных в сфере культуры: проблема перевода качественного контекста в числовой эквивалент; влияние позиции

наблюдателя исследователя на интерпретацию числовых данных в формулировании выводов.

Тема 5. Основы визуализации и программные средства визуализации данных

Основные принципы визуализации: графическое подсвечивание значимых показателей; снижение уровня информационного пресыщения; очевидность внутренних связей между информационными блоками; создание визуальных образов неструктурированной информации.

Способы визуализации данных: отношения; распределение; композиция; сравнение; непрерывные и дискретные числовые зависимости; временные зависимости; географические кластеры; логические структуры.

Инструменты визуализации данных: графики и диаграммы; блок-схемы; таблицы и матрицы; инфографика.

Программные средства визуализации данных: средства создания интерактивных дашбордов и инфографики; визуализация с эффектами анимации; облачные сервисы визуализации данных.

Тема 6. Представление и визуализация данных в сфере культуры

Принципы представления данных в сфере культуры: эффективность восприятия; очевидность новизны материала; иерархия уровней понимания информации; простота передачи сложных идей и закономерностей; информативность; кратчайший путь демонстрации выводов; эстетика.

Целеполагание визуализации и представления данных в сфере культуры: наглядность представления данных, объяснимость идеи, доступность выводов; демонстрация сравнения, организованности, связности данных; привлечение внимания, содействие пониманию, побуждение.

Базовые этапы и правила представления данных в сфере культуры. Виды инфографики представления данных: функциональные, системно-логические образно-ассоциативные.

Программные средства и онлайн сервисы создания презентаций и инфографики. Средства и способы демонстрации, размещения и продвижения результатов культурологических исследований, основанных на анализе больших данных в интернет-пространстве.

ПРИМЕРНАЯ УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА

очная форма получения высшего образования

Разделы и темы	Количество аудиторных часов					Количество часов УСР	Форма контроля
	всего	лекции	лабораторные занятия	практические занятия	семинары		
Тема 1. Большие данные и интелектуальный анализ данных в сфере культуры	6	2		2	2		
Тема 2. Методы и подходы к анализу больших данных	8	2		4	2	12	отчет
Тема 3. Программные средства для анализа данных	8	2		4	2	12	отчет
Тема 4. Основы интерпретации результатов анализа данных и специфика в сфере культуры	6	2		2	2	5	презентация
Тема 5. Основы визуализации и программные средства визуализации данных	8	2		4	2	12	отчет
Тема 6. Представление и визуализация данных в сфере культуры	8	2		2	4	5	презентация
Всего аудиторных	**44**	**12**		**18**	**14**	**46**	**зачет**
Всего	**90**						

ПРИМЕРНАЯ УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА

заочная форма получения высшего образования

Разделы и темы	Количество аудиторных часов					Количество часов УСР	Форма контроля
	всего	лекции	лабораторные занятия	практические занятия	семинары		
Тема 1. Большие данные и интелектуальный анализ данных в сфере культуры	1	1					
Тема 2. Методы и подходы к анализу больших данных	2	1			1	18	отчет
Тема 3. Программные средства для анализа данных	2			2		20	отчет
Тема 4. Основы интерпретации результатов анализа данных и специфика в сфере культуры	2	1			1	12	презентация
Тема 5. Основы визуализации и программные средства визуализации данных	2	1		1		18	отчет
Тема 6. Представление и визуализация данных в сфере культуры	1			1		12	презентация
Всего аудиторных	**10**	**4**		**4**	**2**	**80**	**зачет**
Всего	**90**						

РАЗДЕЛ КОНТРОЛЯ ЗНАНИЙ

Вопросы к зачёту

1. Определение больших данных (Big Data)
2. Современные подходы к обработке и хранению больших данных
3. Особенности Больших данных сферы культуры
4. Интеллектуальный анализ больших данных (Data Mining) и области применения
5. Задачи интеллектуального анализа больших данных: прогнозирование и предсказание
6. Задачи интеллектуального анализа больших данных: классификация и сегментация
7. Задачи интеллектуального анализа больших данных: выявление скрытых связей и закономерностей
8. Этапы интеллектуального анализа: предварительный анализ данных и
9. применение алгоритмов машинной обработки данных
10. Этапы интеллектуального анализа: интерпретация результатов и применение результатов анализа для принятия решений
11. Подходы к изучению культурных феноменов и процессов цифровой среды
12. Основные понятия статистики: шкалы измерений; генеральная совокупность и выборка; случайная величина и ее распределение. Статистическая гипотеза
13. Математические методы и компьютерные технологи анализа данных в сфере культуры: статистические методы,
14. Математические методы и компьютерные технологи анализа данных в сфере культуры: математическое моделирование,
15. Математические методы и компьютерные технологи анализа данных в сфере культуры: глубинный анализ данных и машинное обучение
16. Математические методы и компьютерные технологи анализа данных в сфере культуры: сетевой анализ и анализ сложных динамических систем.
17. Описательная статистика.
18. Статистический вывод. Уровень достоверности. Статистические критерии.
19. Дисперсионный, корреляционный и регрессионный анализ.
20. Машинное обучение. Алгоритмы, обучающиеся по данным.
21. Машинное обучение с «учителем» и «без учителя»

22. Кластерный анализ. Основные цели кластеризации. Типы кластерных структур

23. Использование искусственного интеллекта для сбора, извлечения и анализа данных

24. Интерпретация результатов анализа данных: объяснение и обобщение

25. Интерпретация результатов анализа данных: теоретическая обработка данных vs эмпирическая обработка данных

26. Интерпретация результатов анализа данных: преобразование данных результата статистического анализа в эмпирические знания

27. Интерпретация результатов анализа данных: получение теоретических знаний на базе эмпирических знаний

28. Теоретические аспекты интерпретации результатов анализа данных: проверка, уточнение, модификация исходных гипотез; определение отношений между данными и гипотезами

29. Теоретические аспекты интерпретации результатов анализа данных: развитие гипотез до уровня теоретических высказываний; объяснение проблем и их решение

30. Этапы интерпретации результатов анализа данных: структурирование данных; сопоставление результатов с данными из других источников; проведение многоаспектного анализа

31. Этапы интерпретации результатов анализа данных: определение причинно-следственных связей; обобщение и синтез всей полученной информации; формулирование заключительных выводов

32. Специфика интерпретации данных в сфере культуры: проблема перевода качественного контекста в числовой эквивалент; влияние позиции наблюдателя исследователя на интерпретацию числовых данных в формулировании выводов

33. Принципы представления данных в сфере культуры: эффективность восприятия и очевидность новизны материала

34. Принципы представления данных в сфере культуры: иерархия уровней понимания информации, простота передачи сложных идей и закономерностей.

35. Принципы представления данных в сфере культуры: информативность и кратчайший путь демонстрации выводов, эстетика.

36. Целеполагание визуализации и представления данных в сфере культуры: наглядность представления данных и объяснимость идеи,

37. Целеполагание визуализации и представления данных в сфере культуры: побуждение и привлечение внимания

38. Целеполагание визуализации и представления данных в сфере культуры: доступность выводов и содействие пониманию

39. Целеполагание визуализации и представления данных в сфере культуры: демонстрация сравнения, организованности, связности данных

40. Базовые этапы и правила представления данных в сфере культуры

В качестве формы промежуточной аттестации рекомендован устный опрос.

РЕКОМЕНДУЕМЫЕ МЕТОДЫ ПРЕПОДАВАНИЯ

В процессе преподавания дисциплины используются эффективные педагогические методы и технологии: проблемно-ориентированная технология обучения; коммуникативные и информационные технологии; технологии учебной и исследовательской деятельности; метод анализа конкретных ситуаций, другие методики.

Методические рекомендации по организации и выполнению самостоятельной работы студентов

Для эффективного освоения студентами дисциплины «Анализ данных и визуализация в культуре» используются педагогические методики и технологии, способствующие приобщению студентов к поисковой работе, технологии учебно-исследовательской деятельности, коммуникативные технологии (дискуссии, учебные дебаты и др.), игровые технологии (деловые, ролевые, имитационные игры) и др.

Самостоятельная работа студентов является основным способом охвата учебного материала по дисциплине «Анализ данных и визуализация в культуре» в свободное от обязательных учебных занятий время. Цель самостоятельной работы студентов – содействие усвоению в полном объеме содержания учебной дисциплины через систематизацию, планирование и контроль собственной деятельности.

При организации самостоятельной работы студентов необходимо придерживаться следующих рекомендаций:

– информирование студентов с первой недели семестра об учебных заданиях на самостоятельную проработку отдельных тем или их частей, семинарских и практических занятий с последующим контролем их выполнения;

– проработка обзорного лекционного материала, изучение по учебным пособиям программного материала и рекомендованных преподавателем литературных источников;

– контент-анализ публикаций по анализу данных и визуализации в культуре, составление аннотаций и реферирование;

– организация самостоятельной работы студентов в форме делового взаимодействия, когда студент получает конкретные указания и рекомендации об организации и содержания самостоятельной деятельности и преподаватель выполняет функцию управления через контроль и коррекцию ошибочных действий;

– разработка тематических презентаций;

– текущий контроль самостоятельной работы студента в виде тестирования, проверки выполнения заданий по УСР, конспектов и др.

Перечень используемых средств диагностики результатов учебной деятельности по дисциплине «Анализ данных и визуализация в культуре»

Оценка учебных достижений студента осуществляется с использованием фонда оценочных средств и технологий УВО. Фонд оценочных средств учебных достижений студента включает:

– типовые задания в различных формах (устные, письменные, тестовые, ситуационные и т.п.);

– учебно-исследовательская работа студентов;

– иные средства диагностики в соответствии с учебной программой.

Фонд технологий контроля обучения включает:

– устный опрос во время семинарских занятий;

– выступление студентов на семинарских и практических занятиях с разработанными ими темами и заданиями;

– подготовка презентаций;

– аттестацию по окончании изучения дисциплины с применением устной и письменной методик контроля обучения.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Литература

Основная

1. Комалова, Л. Р. Современная информационная среда и наукометрия : учебное пособие / Л. Р. Комалова. – Москва : Проспект, 2021. – 104 с.
2. Скакун, В. В. Системы управления базами данных : учебно-методическое пособие для студентов учреждений высшего образования по специальности 1-98 01 01 "Компьютерная безопасность (по направлениям)", направление специальности 1-98 01 01-02 "Компьютерная безопасность (радиофизические методы и программно-технические средства)" / В. В. Скакун. - Минск : БГУ, 2020. – 158 с.
3. Информатика для гуманитариев : учебник и практикум для студентов высших учебных заведений, обучающихся по гуманитарным направлениям и специальностям / под ред. Г. Е. Кедровой. - 2-е изд. - Москва : Юрайт, 2021. - 653 с.
4. Федотова, Е. Л. Прикладные информационные технологии : учебное пособие для студентов, обучающихся по профилю "Инфоромационный менеджмент" направления 38.03.02 "Менеджмент" / Е. Л. Федотова, Е. М. Портнов. - Москва : ФОРУМ-ИНФРА-М, 2020. - 334 с.
5. Мясоедов, С. П. Кросс-культурный менеджмент : учебник для студентов высших учебных заведений, обучающихся по экономическим направлениям / С. П. Мясоедов, Л. Г. Борисова. - 3-е изд. - Москва : Юрайт, 2021. - 313 с.
6. Борзова, Е. П. Сравнительная культурология : учебник для студентов высших учебных заведений, обучающихся по гуманитарным направлениям / Е. П. Борзова. - 2-е изд., перераб. и доп. - Москва : Юрайт, 2021. – 554 с.