

ИССЛЕДОВАНИЕ ДАННЫХ

Петрова С.П.

(научный руководитель – Ерохина Ю.А.)

Белорусский национальный технический университет

Минск, Беларусь

Аннотация. В настоящей работе рассматривается понятие «исследование данных» и его применение в различных сферах деятельности человека. Показано, как именно происходит процесс исследования и какие результаты можно получить при их использовании.

Введение

Под исследованием данных понимают комплексный процесс, позволяющий не просто собирать информацию, а превращать ее в полезные знания, которые помогают решать реальные проблемы и принимать эффективные решения.

Историческая справка

В качестве основоположников развития теории исследования данных можно выделить следующих ученых и их научные работы:

Дэвид Бейкер является создателем первой системы сбора данных, которые использовались в медицинских целях для диагностики болезней и прогнозирования осложнений. Основной идеей являлось использование вычислительных методов для разработки и предсказания структуры и функций белков, а также для создания новых белков с заданными свойствами [1].

Джеймс Саймонс – основатель компании «Renaissance Technologies», которая известна своей успешной торговой стратегией, основанной на математических моделях, исследовании и анализа данных. Предложил собирать и анализировать огромные объемы данных из различных источников (финансовые рынки, новости, социальные сети и т.д.), чтобы выявить статистические закономерности и аномалии, которые могут быть использованы для прогнозирования цен [2].

Крис Андерсон – автор книги «Длинный хвост», в которой доказал важность анализа данных для понимания потребительских предпочтений и формирования стратегии маркетинга. Концепция данной книги повлияла на индустрию развлечений, способствуя развитию стриминговых сервисов, радио и платформ, предлагающих широкий выбор контента [3].

Этапы исследования данных

1. Сбор данных. Первоначально определяется источники данных и сбор этих данных. В качестве источников могут выступать базы данных, файлы, сенсоры, социальные сети, веб-сайты и др.

2. Очистка данных. Собранные данные часто содержат ошибки, погрешности, пропуски и несоответствия. Под очисткой данных понимают процесс устранения таких проблем, что позволяет получить надежные и точные данные для их дальнейшего анализа.

3. Анализ данных включает в себя использование различных методов и технологий для изучения данных и выявления скрытых закономерностей и тенденций. В зависимости от целей исследования, используются различные методы статистического анализа, машинного обучения, визуализации данных и другие методы.

4. Интерпретация результатов. На основе результатов анализа данных делаются выводы и формулируются рекомендации. Интерпретация результатов анализа должна обеспечить доступность и однозначность понимания для всего круга заинтересованных лиц, вне зависимости от их специализированных знаний.

Методы исследования данных

Задачи, возникающие при анализе информации, достаточно разнообразны. В связи с этим качество методов исследования данных выделяют следующие.

Статистический анализ – это способ проверки программного кода без его фактического запуска.

Машинное обучение – использует алгоритмы для обучения компьютера на основе данных и построения моделей, которые могут предугадывать будущие события или результаты. Есть несколько видов машинного обучения: ансамблевые методы, обучение с подкреплением, классическое и глубокое обучение [4].

Визуализация данных – представляет данные в виде графиков, диаграмм, карт и других визуальных изображений, что позволяет лучше понять структуру данных и выявлять скрытые тенденции.

Корреляционный анализ – анализ, который показывает насколько сильно изменения в одной переменной связаны с изменениями в другой переменной.

Примеры исследования данных в разных сферах

В настоящее время интерес к исследованию данных возрастает в самых разных областях человеческой деятельности. В качестве примера можно привести следующее [4]:

- бизнес: анализ покупательских предпочтений, оптимизация маркетинговых кампаний, прогнозирование продаж;
- наука, в частности медицина: изучение климата, моделирование экосистем, открытие новых лекарств, диагностика болезней, разработка новых методов лечения, прогнозирование осложнений;
- социальные сети: анализ поведения пользователей, идентификация трендов, борьба с дезинформацией;
- производство: разработка системы, которая предсказывает вероятность поломки оборудования в ближайшем будущем, позволяя проводить профилактическое обслуживание и предотвратить дорогостоящие простои;
- транспорт и логистика: разработка наиболее эффективных маршрутов доставки, сокращение времени и затрат на транспортировку;
- образование: разработка для каждого студента индивидуальной программы обучения, соответствующую его потребностям и способностям;
- энергетика: разработка модели, которая с высокой точностью предсказывает выработку электроэнергии, позволяя эффективно планировать распределение электроэнергии и т.д.

Ответственность и этические аспекты исследования данных

В вопросах исследования данных возникают моменты, связанные с этическими аспектами [5].

1. Важно учитывать конфиденциальность личной информации и защищать данные от несанкционированного доступа.
2. Необходимо обеспечивать справедливое и непредвзятое использование данных.
3. Важно обеспечить прозрачность и подотчетность в процессе исследования данных.

Заключение

Исследование данных – это мощный инструмент, который позволяет получать ценную информацию из данных и применять ее для решения реальных проблем. В будущем исследование данных будет играть еще более важную роль в развитии здравоохранения (прогнозирование эпидемий и оптимизация распределения ресурсов), транспорта (автономное вождение, прогнозирование трафика), финансах (оценка кредитного риска, обнаружение мошеннических транзакций) и др.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Нобелевскую премию по химии присудили за изучение структуры белков, Интерфакс: В мире [Электронный ресурс]. – Режим доступа: <https://www.interfax.ru/world/985929>. – Дата доступа: 08.04.2025;

2. Саймонс Джеймс Харри-Рувики: Интернет-энциклопедия [Электронный ресурс]. – Режим доступа:
https://ru.ruwiki.ru/wiki/%D0%A1%D0%B0%D0%B9%D0%BC%D0%BE%D0%BD%D1%81_%D0%94%D0%B6%D0%B5%D0%B9%D0%BC%D1%81_%D0%A5%D0%B0%D1%80%D1%80%D0%B8%D1%81. – Дата доступа: 08.04.2025;
3. Крис Андерсон: «Длинный хвост» технологий, Идеономика – Умные о главном [Электронный ресурс]. – Режим доступа: <https://ideanomics.ru/lectures/14759>. – Дата доступа: 08.04.2025;
4. РБК Тренды – кто и как меняет глобальную экономику [Электронный ресурс]. – Режим доступа: <https://trends.rbc.ru/trends/amp/news/60c85c599a7947f5776ad409>. – Дата доступа: 08.04.2025;
5. Этика обработки данных и конфиденциальность при анализе: Обеспечение ответственного использования данных [Электронный ресурс]. – Режим доступа: <https://falconediting.com/ru/blog/etika-obrabotka-dannykh-i-konfidetsialnost-pri-analize-obespechenie-otvetstvennogo>. – Дата доступа: 09.04.2025.