Лекция 2. Источники и типы данных

(структурированные, неструктурированные, большие данные; этапы сбора и хранения)

---

1. Введение

Современное общество производит колоссальные объёмы данных. Каждое действие в цифровом пространстве — сообщение, покупка, поиск, регистрация или даже движение устройства — оставляет цифровой след. Данные стали стратегическим ресурсом XXI века, их называют «новой нефтью» цифровой экономики. Однако ценность данных проявляется только тогда, когда они собраны, организованы и проанализированы.

Для того чтобы эффективно применять методы интеллектуального анализа данных и машинного обучения, необходимо понимать, **откуда берутся данные, какие они бывают, и как организуется их хранение и обработка.**

---

2. Источники данных

Источники данных могут быть самыми разнообразными — от традиционных корпоративных баз данных до потоков информации из интернета вещей. Основные типы источников включают:

1. **Внутренние источники организации**
    - Базы данных предприятий (CRM, ERP, бухгалтерия, HR-системы);
    - Лог-файлы и отчёты;
    - История транзакций, обращения клиентов, электронные письма;
    - Производственные датчики и системы мониторинга.
2. **Внешние источники**
    - Открытые данные (open data) — государственные порталы, статистические агентства, международные организации;
    - Социальные сети и интернет-платформы (Facebook, X, Instagram, YouTube, Telegram);
    - Веб-сайты, новостные порталы, блоги;
    - Маркетинговые и коммерческие базы (например, отчёты компаний, демографические данные);
    - API и онлайн-сервисы (например, Google Maps, Twitter API).
3. **Интернет вещей (IoT)**
    - Датчики, счётчики, устройства «умного дома», транспортные системы, промышленное оборудование;

    - Такие устройства генерируют миллионы записей в реальном времени.
4. **Экспериментальные и исследовательские данные**
    - Научные измерения, лабораторные эксперименты, медицинские обследования;
    - Биометрические и генетические данные.
5. **Мультимедийные источники**
    - Фото, видео, аудио, сканированные документы, спутниковые изображения;
    - Часто используются в анализе изображений и распознавании речи.

---

3. Типы данных

Данные могут классифицироваться по степени структурированности и форме представления.

3.1. Структурированные данные

Это данные, организованные по чёткой схеме, чаще всего в виде таблиц. Каждый элемент данных имеет определённый тип (число, строка, дата и т.п.), а строки и столбцы соответствуют объектам и их признакам.

Примеры:

- Таблицы продаж, банковские счета, клиентские базы;
- Реляционные базы данных (MySQL, Oracle, PostgreSQL).

Преимущества:

- Удобство хранения и обработки;
- Возможность быстрого поиска и фильтрации;
- Простота интеграции с аналитическими системами.

Недостатки:

- Ограниченная гибкость (сложно хранить мультимедийные данные);
- Не подходит для нестандартных или постоянно меняющихся структур.

3.2. Неструктурированные данные

Это информация, не имеющая фиксированного формата или структуры. Примеры — тексты, изображения, видео, аудиозаписи, электронные письма, посты в социальных сетях.

Особенности:

- 80–90% всех данных в мире являются неструктурированными;
- Их обработка требует применения технологий искусственного интеллекта (NLP, компьютерное зрение).

Примеры использования:

- Анализ отзывов и комментариев (Sentiment Analysis);
- Распознавание лиц и объектов на изображениях;
- Анализ голосовых сообщений.

Хранилища:
Для таких данных применяются NoSQL-системы — MongoDB, Cassandra, ElasticSearch, Hadoop HDFS.

3.3. Полуструктурированные данные

Это промежуточная форма между структурированными и неструктурированными данными.
Они содержат теги или атрибуты, указывающие структуру, но при этом сохраняют гибкость.

Примеры:

- XML, JSON, HTML-документы;
- Логи серверов и сетевых устройств.

Преимущества:

- Гибкость и возможность хранения сложных объектов;
- Легкость передачи между системами (например, в веб-приложениях).

3.4. Большие данные (Big Data)

Большие данные — это совокупность объёмных, быстро поступающих и разнообразных данных, которые невозможно обработать традиционными методами.

Обычно их описывают через концепцию **5V**:

1. **Volume (объём)** — терабайты и петабайты информации;
2. **Velocity (скорость)** — данные поступают в реальном времени;
3. **Variety (разнообразие)** — структурированные, неструктурированные и полуструктурированные данные;
4. **Veracity (достоверность)** — необходимость фильтрации ошибок и ложных сведений;
5. **Value (ценность)** — получение практической пользы и знаний.

Примеры больших данных:

- Потоки кликов на сайте;
- Телеметрия автомобилей;
- Данные со спутников или смарт-устройств;
- Медицинские сенсоры.

Инфраструктура Big Data:

- Hadoop, Spark, Hive, Kafka, Google BigQuery, Amazon S3.

---

4. Этапы сбора данных

Сбор данных — это первый шаг в процессе анализа. Он включает несколько последовательных этапов:

1. **Определение целей сбора**
   — какие данные нужны и для какой задачи.
2. **Выбор источников**
   — внутренние системы, внешние API, открытые наборы данных.
3. **Извлечение данных (Data Extraction)**
   — использование SQL-запросов, парсинг сайтов, автоматизированные сенсоры.
4. **Очистка данных (Data Cleaning)**
   — удаление дубликатов, исправление ошибок, обработка пропусков.
5. **Интеграция и преобразование**
   — объединение данных из разных форматов и систем, стандартизация.
6. **Загрузка в хранилище (Data Loading)**
   — сохранение в базы данных, озёра данных (Data Lake), облачные хранилища.

---

5. Хранение данных

После сбора данные должны быть надёжно сохранены, структурированы и защищены.

Основные подходы:

1. **Реляционные базы данных (SQL)**
   Используются для структурированных данных (PostgreSQL, MySQL, Oracle).
2. **NoSQL-хранилища**
   Подходят для больших и гибких массивов (MongoDB, Cassandra, Redis).
3. **Data Warehouse (Хранилище данных)**
   Централизованное хранилище для аналитики, объединяющее данные из разных источников.
4. **Data Lake (Озеро данных)**
   Хранилище в «сыром» виде — для последующей обработки (Hadoop, Amazon S3, Azure Data Lake).
5. **Облачные технологии**
   Используются для масштабируемости и снижения затрат (Google Cloud, AWS, Microsoft Azure).

---

6. Безопасность и качество данных

Для надёжной аналитики важно обеспечить:

- Защиту данных от несанкционированного доступа;
- Контроль целостности и резервное копирование;
- Проверку достоверности источников;
- Управление жизненным циклом данных (Data Governance).

---

7. Заключение

Понимание источников и типов данных является фундаментом для эффективной аналитики и машинного обучения.
Данные — это не просто цифры и тексты, а стратегический актив, требующий грамотного управления.
От того, как данные собираются, хранятся и обрабатываются, напрямую зависит качество будущих моделей и решений.

Список литературы

1. Хэн, Дж., Камбер, М., Пей, Дж. *Интеллектуальный анализ данных: концепции и методы*. — М.: Вильямс, 2019.
2. Provost, F., Fawcett, T. *Data Science for Business*. — O’Reilly Media, 2013.
3. Marr, B. *Big Data in Practice*. — Wiley, 2016.
4. Russom, P. *Big Data Analytics*. — TDWI Research, 2019.
5. Ларсон, Д., Шин, А. *Основы Data Science*. — СПб.: Питер, 2021.