Лекция. Жизненный цикл данных

8 марта 2023 г.

ВОПРОСЫ:

- Классификация данных по степени структурированности
- Большие данные
- Характеристики Big Data
- Источники Big Data

1 Классификация данных по степени структурированности

По степени структурированности можно выделить:

структурированные данные — данные, имеющие строго фиксированную структуру, определяемую формальной моделью данных (например, реляционной схемой);

полуструктурированные (слабоструктурированные) данные — данные, не имеющие строго определенной структуры, но предполагающие наличие правил, позволяющих выделять отдельные семантические элементы при их интерпретации, прежде всего правил расстановки тегов и других маркеров, отмечающих и выделяющих элементы данных (например, файлы, созданные с использованием языка XML и его многочисленных производных, html-страницы и др.); неструктурированные данные — данные, произвольные по форме, не имеющие строго определенной структуры и не организованные по определенным правилам.

К структурированным данным относятся данные, определяющие конкретную предметную область. Такие данные упорядочены специальным образом и организованы таким образом, чтобы над такими данными можно было выполнить анализ. Обычно такие данные хранятся в виде таблиц в реляционных базах данных. Почти все алгоритмы машинного обучения и Data Mining работают со структурированными данными. К полуструктурированным данным относятся данные, которые не соответствуют чёткой структуре таблиц и отношений в реляционных базах данных, однако такие данные

содержат специальные теги и иные маркёры, позволяющие отделить семантические элементы. Такие данные принадлежат одному классу, но при этом могут иметь разные атрибуты. Xml документы являются простейшем примером полуструктурированных данных. К неструктурированными данным относятся данные, которые не имеют определённой формы, могут включать в себя видео, аудио файлы, свободный текст, информацию, поступающую из социальных сетей. На сегодняшний день 80 процентов информации входит в группу неструктурированной. Такую информацию необходимо комплексно анализировать, для упрощения ее дальнейшей обработки.

Отдельно следует определить **машинные данные и потоковые данные**. К машинным данным относится информация, автоматически генерируемая компьютером, процессом, приложением или устройством без вмешательства человека (когда мы говорим об интернете вещей или о данных медицинского мониторинга, имеются в виду как раз машинные данные). Машинные данные становятся одним из основных источников информации, это в первую очередь относится к данным контроля и аудита (то есть к сведениям, фиксируемым в различных журналах регистрации).

Потоковые данные могут относиться почти к любой из перечисленных выше категорий, однако у них имеется одно дополнительное свойство. Данные поступают в систему при возникновении некоторых событий, а не загружаются в хранилище данных большими массивами. Примером потоковой обработки данных является сервис YouTube, проводящий анализ данных пользователей исходя не только из просмотренных ими полностью видеозаписей и трансляций, но и из материалов, которые они пропустили. Другим примером могут служить данные телеметрии, полученные с любого датчика или набора датчиков (например, системы «умный дом»).

Операционные данные — это данные о клиентах, поставщиках, партнерах и сотрудниках, доступные в процессе онлайн-обработки транзакций и/или полученные из онлайн-базы данных аналитической обработки. Чаще всего их собирают с помощью датчиков и мониторинга процессов предприятий. Их источником могут быть, например, кассовые аппараты, подключенные к банковской системе, интеллектуальные счетчики, голосовая связь. «Темные данные» организации не собирают или не хранят специально; они формируются (попутно) в процессе ведения бизнеса или взаимодействия с сетевыми сервисами, после чего остаются в интернет-архивах. К ним относятся электронные письма, мультимедиа, системные журналы. Публичные данные распространяются госорганами и коммерческими компаниями. Их ценность раскрывается в совокупности с другими источниками; они позволяют определить направления социально-экономического развития в отдельном городе, в стране или в группе стран.

2 Большие данные)

Это информация, которую уже невозможно обрабатывать традиционными способами, в том числе структурированные данные, медиа и случайные

объекты

Большие данные – комплексный набор методов, подходов и инструментов обработки структурированных и неструктурированных данных колоссальных объемов. Главной целью обработки Big Data является быстрое и эффективное использование всех видов информации в условиях непрерывного изменения и прироста в больших объёмах. **Big data** представляет собой безмерный объем информации, который не может быть обработан стандартными инструментами и аппаратными средствами. Основными задачами Big Data являются хранение и обработка информации гигантских объёмов данных. Большие данные по сравнению с обычными данными требуют иной подход к обработке. При обработке Big data используются собственные инструменты и технологии, которые предназначены для данных со сверхбольшим объёмом информации. **Термин Big Data** был предложен Клиффордом Линчем (Clifford Lynch), редактором журнала Nature, который 3 сентября 2008 года выпустил отдельный номер, главной темой которого была «Как могут повлиять на будущее науки, технологии, открывающие возможности работы с большими объёмами данных?» (Оригинал: «Big data: How do your data grow?»). Термин «Большие данные» был предложен по аналогии с терминами «Большая нефть», «Большая руда» и т. д. Размер больших данных в 2012 году определялся от нескольких десятков терабайт до петабайт. Термин большие данные может быть причислен к данным, связанным с высочайшей изменчивостью источников данных, а также обладающим сложными взаимосвязями и трудностями изменения или удаления отдельных записей. Большие данные характеризуются гигантским объёмом, значительной скоростью поступления данных, а также многообразием самих данных. Для таких данных требуются новейшие способы обработки, которая в дальнейшем может привести к улучшению методов принятия решений, оптимизации процессов и поиска закономерностей.

К 2011 году понятие Big Data стало набирать популярность, в основном, в крупных корпорациях таких как Microsoft, IBM, Oracle, EMC, HP и др. В 2011 году исследовательская компания Gartner отмечает большие данные как тренд номер два в информационно-технологической инфраструктуре после виртуализации

3 Характеристики Big Data

Существует множество характеристик для Больших данных, но здесь будут рассматриваться самые основные. Сфера Больших Данных характеризуется следующими признаками: **Volume (объем)**: накопленная база данных представляет собой гигантский объем информации, для которого обработка и хранение традиционными способами являются трудоёмкими процессами. Такой объем нуждается в новых подходах и в более усовершенствованных инструментах.

Velocity (скорость): данный признак указывает как на увеличивающуюся скорость накопления, так и на скорость обработки данных. В последнее

время стали более востребованы технологии обработки данных в реальном времени.

Variety (многообразие): данная характеристика означает возможность одновременной обработки структурированной и неструктурированной информации различных форматов. Главным отличием структурированной информации является возможность классификации. Примером такой информации может служить информация о клиентских транзакциях.

Veracity (достоверность данных): в настоящее время достоверность имеющихся данных является важнейшим критерием для пользователей. Недостоверная информация приводит к затруднению анализа данных. Value (ценность накопленной информации): Большие Данные должны быть полезны в усовершенствовании бизнес-процессов, составлении отчетности или оптимизации расходов компаний. Первые три характеристики определяют так называемый принцип «Трёх V» Объём относится к наборам данных, размер которых выходит за пределы возможностей программных средств типичной базы данных сбора, хранения, обработки и анализа данных. Разнообразие определяет способность обработки множества типов, источников и форматов данных от сенсоров, умных устройств, социальных сетей. Также разнообразие характеризуется способностью интегрировать все большее число источников, содержащих различные структурированные, полуструктурированные данные, извлекаемыми из web-страниц, web log файлов, e-mail, документов и др. Скорость определяет реакцию на текущую информацию за время, ограниченное приложением. Примером является потоковая обработка (например, GPS данных в реальном времени)

4 Источники Big Data

- Социальные сети и их данные
- Данные от измерительных устройств
- Журналы доступа пользователей веб-сайтов
- Сенсорные сети
- Тексты и документы из Интернета
- Научные данные (астрономия, геном человека, исследования атмосферы, биохимия, биология)
- Данные министерства обороны
- Медицинские наблюдения
- Фото- и видео-архивы
- Данные электронной коммерции

Основной причиной появления больших данных являются достижения в области мобильных устройств, такие как цифровое видео, фотографии, аудио, а также современные системы электронной почты и обмена текстовыми сообщениями. Пользователи получают данные в количествах, которые нельзя было представить десять лет назад; при этом появляются новые приложения, такие как Google Translate, предоставляющие функции сервера больших данных – перевод произнесенных или введенных с мобильных устройств фраз. Корпорация IBM в отчете "Тенденции развития технологий в 2013 году говорит в первую очередь о доступе к большим данным с мобильных устройств и характеризует большие данные по объему (volume), разнообразию (variety), скорости (velocity) и достоверности (veracity). Эти данные гораздо менее структурированы, чем записи в реляционных базах данных, но могут быть прокоррелированы с ними.

5 Наборы данных для машинного обучения

- Google Dataset Search
- Kaggle
- UCI Machine Learning Repository
- VisualData
- Find Datasets

Государственные датасеты

- Data.gov. Здесь можно найти данные от разных государственных учреждений США. Они варьируются от государственных бюджетов до школьных оценок.
- Food Environment Atlas. Содержит данные о том, как различные факторы (близость магазинов/ресторанов, цены на продукты и тому подобное) влияют на выбор продуктов и качество питания в США.
- School system finances. Данные о финансах школьных систем в США.
- Chronic disease data. Данные о показателях хронических заболеваний на территории США.
- The US National Center for Education Statistics. Данные об образовательных учреждениях и образовательной демографии в США и во всём мире.
- The UK Data Service. Крупнейшая в Великобритании коллекция социальных, экономических и демографических данных.
- Data USA

Экономика и финансы

- Quandl. Хороший источник экономических и финансовых данных — полезен при построении моделей для прогнозирования экономических показателей или цен на акции.
- World Bank Open Data. Наборы данных, охватывающих демографическую ситуацию, огромное количество экономических показателей и индикаторов развития со всего мира.
- IMF Data. Международный валютный фонд публикует данные о международных финансах, показателях долга, валютных резервах, инвестициях и ценах на сырьевые товары.
- Financial Times Market Data. Актуальная информация о финансовых рынках со всего мира, которая включает индексы цен на акции, товары и валюту.
- Google Trends. Изучайте и анализируйте данные о поисковой активности в Интернете и трендах по всему миру.

merican Economic Association (AEA). Хороший источник данных о макроэкономике США.

Компьютерное зрение

- xView. Один из самых больших общедоступных наборов воздушных снимков земли. Он содержит изображения различных сцен со всего мира, аннотированных с помощью ограничительных рамок.
- Labelme. Большой датасет аннотированных изображений.
- ImageNet. Датасет изображений для новых алгоритмов, организованный в соответствии с иерархией WordNet, в которой сотни и тысячи изображений представляют каждый узел иерархии.
- LSUN. Датасет изображений, разбитых по сценам и категориям с частичной разметкой данных.
- MS COCO. Крупномасштабный датасет для обнаружения и сегментации объектов.
- COIL100. 100 разных объектов, изображённых под каждым углом в круговом обороте.
- Visual Genome. Датасет с 100 тыс. подробно аннотированных изображений.
- Google's Open Images. Коллекция из 9 миллионов URL-адресов к изображениям, «которые были помечены метками, охватывающими более 6000 категорий» под лицензией Creative Commons.

- Labelled Faces in the Wild. Набор из 13 000 размеченных изображений лиц людей для использования приложений, которые предполагают использование технологии распознавания лиц.
- Stanford Dogs Dataset. Содержит 20 580 изображений из 120 пород собак.
- Indoor Scene Recognition. Датасет для распознавания интерьера зданий. Содержит 15 620 изображений и 67 категорий.

Обработка естественного языка

- HotspotQA Dataset. Датасет с вопросами-ответами, позволяющий создавать системы для ответов на вопросы более понятным способом.
- Enron Dataset. Данные электронной почты от высшего руководства Enron.
- Amazon Reviews. Содержит около 35 млн отзывов с Amazon за 18 лет. Данные включают информацию о продукте и пользователе, оценки и сам текст отзыва.
- Google Books Ngrams. Коллекция слов из Google Книги.
- Blogger Corpus. Коллекция из 681 288 постов с Blogger. Каждый блог содержит как минимум 200 вхождений часто используемых английских слов.
- Gutenberg eBooks List. Аннотированный список электронных книг проекта «Гутенберг».
- Hansards text chunks of Canadian Parliament. Датасет с 1.3 миллионами пар текстовых файлов, записанных с дебатов 36-го Канадского Парламента.
- Jeopardy. Архив с более чем 200 000 вопросов с телевикторины Jeopardy.
- Rotten Tomatoes Reviews. Архив из более чем 480 000 рецензий с Rotten Tomatoes.
- SMS Spam Collection in English. Датасет, состоящий из 5574 спам-смс на английском.
- Yelp Reviews. Датасет от Yelp, содержащий более 5 млн отзывов.
- UCI's Spambase. Большой датасет спам-писем.

Медицинские данные MIMIC-III. Датасет с обезличенными данными о состоянии здоровья 40 000 пациентов, находящихся на интенсивной терапии. Он включает демографические данные, показатели жизнедеятельности, лабораторные анализы, лекарства и многое другое.