**Большие данные и машинное обучение**

**BIG DATA AND MACHINE LEARNING**

Кудряшов Г.А.

Санкт-Петербургский государственный экономический университет,

г. Санкт-Петербург, Россия

Раскрыты ключевые понятия и описаны основные методы технологии Больших данных. Рассмотрены ее современные практические приложения как в мире, так и в России на примере некоторых крупных компаний.

Ключевые слова: Большие данные, Раскопка данных, Машинное обучение, Система управления базами данных (СУБД).

Were disclosed key concepts and described basic methods of Big Data technology. Its modern practical applications are considered, both in the world and in Russia by the example of some large companies.

Keywords: Big data, Data mining, Machine learning, Database Management System (DBMS).

Большие Данные – это не только огромные объёмы данных, как можно понять из названия, но это ещё и совершенно разнородные данные, хранящиеся в самых разных источниках. Это не только показатели с известным набором реквизитов, для хранения и обработки которых используются хорошо отработанные к настоящему моменту технологии баз данных, но и огромные объёмы неструктурированных текстовых, графических, геопространственных, аудио и видео данных, генерируемых и распространяемых в Интернете. Поэтому Большие Данные – это ещё и специфические технологии их распределённого накопления, хранения и обработки.

В настоящее время термин «Большие данные» используется для обозначения больших массивов постоянно обновляемой разнородной информации, распределённой по множеству мест хранения, а также совокупности технологий, используемых для их накопления, хранения и обработки.

Для характеристики понятия «Большие Данные» применяется правило VVV, по первым буквам слов английского языка: Volume – объём; Velocity – скорость прироста; Variety – вариабельность, разнородность.

Из правила VVV авторы многих публикаций выводят следующие основные принципы построения технологий работы с Большими Данными:

* Горизонтальная масштабируемость. Система обработки больших данных должна обладать свойством расширяемости. В простейшей интерпретации это означает, что необходимые вычислительные мощности системы растут пропорционально росту объёмов поступающих данных.
* Отказоустойчивость. Наличие резервных вычислительных мощностей на случай временного выхода из строя части основных вычислительных мощностей.
* Локальность данных. Основная часть обработки данных осуществляется в местах их хранения, поскольку расходы на передачу данных для их обработки в другом месте могут стать неоправданно большими.

Стоит отметить, что во многих случаях эти требования невыполнимы в реальности. Например, интеллектуальные поисковые системы не придерживаются принципа локальности данных, поскольку в целях индексации web-страниц они копируют слепок их содержимого на свои серверы.

Необходимость развития технологий обработки Больших Данных обусловлена тем, что их автоматизированный анализ может выявлять важные закономерности, получение которых не под силу человеческому мозгу из-за необъятности сопоставляемых данных и присущих им взаимосвязей.

Во многих случаях это действительно так. Например, анализ статистики поисковых запросов Google позволил выявить, что ещё за несколько недель до наступления эпидемии гриппа, в регионе его распространения наблюдается резкое увеличение числа поисковых запросов по его лечению. Этот факт было бы невозможно установить характерными для человека умозрительными выводами. Практическое следствие: анализ поисковых запросов позволяет предвидеть начало эпидемии раньше, чем это сделает официальная медицина.

Есть и множество других примеров полезного эффекта применения технологий обработки больших данных. Поэтому и утверждается, что благодаря технологиям обработки больших данных могут быть сделаны неожиданные открытия, способные сильно изменить все сферы общественной жизни – от государственного управления до производства и коммуникаций. Исходя из этого, к трём V добавили и четвёртое – Value (ценность), указывающее на социально-экономическую значимость проблем обработки больших данных.

При проведении анализа больших данных используется множество самых разных методов исследования. Единой их классификации не существует, поскольку некоторые классы методов имеют настольно общую трактовку, что могут включать несколько методов анализа, считающихся самостоятельными в других классификациях. Далее приводится именно такая нестрогая классификация

Data Mining (раскопка данных). В широком смысле сюда относят практически все методы, применяющиеся для анализа больших данных. Сюда относятся методы классификации, кластеризации, прогнозирования, визуализации, описательная статистика, корреляционно-регрессионный анализ, нейронные сети, методы поиска ассоциаций, эволюционное и генетическое программирование и даже лингвистический анализ текстов.

Пространственный анализ. Набор методов анализа пространственных данных – топологии местности, географических координат, геометрии объектов, применяемый при обработке данных, собираемых геоинформационными системами (ГИС).

Сетевой анализ. Метод выявления взаимосвязей конкретных сообществ. Позволяет выявлять связи людей, компаний и групп в социальных сетях.

Машинное обучение. Направление в информатике, исследующее возможность создания алгоритмов самообучения компьютера на основе анализа эмпирических данных. Не так давно в моду вошёл другой термин – Глубокое обучение (Deep learning), обозначающий группу алгоритмов, позволяющих существенно усовершенствовать методы настройки нейронных сетей на решение конкретных задач.

Краудсорсинг. В простой интерпретации – извлечение знаний из толпы. Набор методик и способов вовлечения в деятельность по обсуждению проблем и выработке решений широкого круга лиц без оформления с ними трудовых отношений.

Распознавание образов. Многочисленные специализированные приложения методов классификации, кластеризации и искусственных нейронных сетей, входящих в состав методов Data Mining. Широко используются в технологиях распознавания текстов, изображений, фотографий, звуков.

Анализ текстов. Широкий класс методов анализа текстов с целью формальной интерпретации их содержания на предмет наличия в них упоминаний конкретных персон, организаций, событий в положительном или отрицательном смысле.

Визуализация. Методы представления результатов обработки данных в виде диаграмм, графиков или иного рода изображений в целях упрощения их понимания и интерпретации человеком.

Представленный перечень современных методов анализа данных не является исчерпывающим и приведён, главным образом, в целях демонстрации огромного разнообразия методов исследования больших данных с целью выявления в них скрытых закономерностей. Кроме того, что существует большое число методов обработки, используются также определенные технологии в обработке данных.

Технологии работы с большими данными строятся исходя из соответствия правилу VVV и рассмотренных выше основным принципов: горизонтальной масштабируемости, отказоустойчивости и локальности данных. В этой связи можно выделить следующие базовые технологии больших данных.

MapReduce – модель распределённой параллельной обработки больших наборов данных, предложенная корпорацией Google. В соответствии с ней общая задача обработки данных разделяется на несколько более простых заданий, параллельно выполняемых несколькими узлами кластера серверов. По мере выполнения заданий промежуточные данные обобщаются и сводятся в конечный результат. Метод Map разбивает данные на подмножества, а метод Reduce обрабатывает выбранные данные и агрегирует их.

NoSQL – класс СУБД, не использующих язык SQL и предназначенных для управления большими плохо структурированными базами данных. Используются для хранения и обработки данных с постоянно изменяющейся структурой, плохо приспособленных для хранения и обработки традиционными реляционными СУБД. Отказ от применения языка SQL позволяет существенно увеличить скорость доступа к данным и снизить сложность самой СУБД.

Hadoop – набор свободно распространяемых программных средств, предназначенных для разработки и выполнения программ, осуществляющих распределённые вычисления в кластерах из большого числа ЭВМ. Является одной из базовых технологий больших данных, реализующих концепцию MapReduce. Система, построенная на технологиях Hadoop, может масштабироваться практически неограниченно. Используется для обеспечения механизмов контекстного поиска многих высоконагруженных сайтов. В настоящее время почти все инструменты анализа больших данных имеют средства интеграции с Hadoop.

R – язык и свободно распространяемая среда программирования статистических вычислений и графической визуализации данных. Имеет огромное число пакетов расширений, реализующих практически все, имеющиеся на текущий момент методы анализа данных. Является фактическим стандартом при проведении статистических исследований и разработки приложений для анализа данных.

Все вышеперечисленные методы и технологии обработки используются в Big data, и потому, если компания или страна заявляет о том, что она использует данную технологию, то это значит лишь то, что какие-либо средства выше были задействованы в обработке их данных.

В этой связи имеет смысл рассмотреть реальные примеры применения технологий больших данных иностранными и российскими компаниями. Все приведённые далее примеры взяты из открытых источников.

Банк HSBC применяет технологии больших данных для противодействия мошенническим операциям с пластиковыми картами. По оценке банка, благодаря применению этих технологий удалось существенно повысить эффективность службы безопасности и выявить неизвестные ранее организованные преступные группы и мошеннические схемы, которые могли бы нанести ущерб в размере не менее $10 млн.

Компания Luxotticagroup – производитель спортивных очков марок Ray-Ban, Persol и Oakley – применяет технологии больших данных для организации персонализированного смс-маркетинга.

В России технологии больших данных пока используют только крупные организации. По данным открытых источников они внедрены в Сбербанке, Газпромбанке, ВТБ24, «Альфа-Банке», «Райффайзенбанке», «Ситибанке», «Нордеа-Банке», «ОТП Банке», компании «Тройка Диалог», у основных телеком-операторов, в розничных сетях X5 Retail Group, «Глория Джинс», «Юлмарт», «Лента», «М. Видео», Ozon, «Азбука вкуса», в нефтяных компаниях - «Роснефть» и «Сургутнефтегаз». По неподтверждённым данным технологии больших данных применяют Федеральная налоговая служба, аналитический центр правительства России, Пенсионный фонд, правительство Москвы, Фонд обязательного медицинского страхования, Федеральная служба безопасности и Служба внешней разведки.

Сбербанк. Технология Big data применяется для управления рисками, противодействия мошенничеству, оценки кредитоспособности и сегментации клиентов, управления персоналом, прогнозирования загруженности отделений, расчёта бонусов сотрудников и других задач. Разрабатывается собственная база биометрических данных клиентов. Предполагается, что система распознавания лиц позволит существенно повысить эффективность мер по предотвращению мошенничества. Проводятся работы по персонализации предложений клиентам на основе анализа собранных о них данных.

Сургутнефтегаз. Платформа SAP HANA используется для автоматизации учёта продукции, расчёта цен, информационного обеспечения сотрудников.

Из рассмотренных примеров следует, что практически все перечисленные компании, так или иначе применяют технологии анализа больших данных в маркетинге. Здесь действительно эти технологии могут дать определённый эффект за счёт персонализации рекламных сообщений, формирования расширенных профилей целевых клиентов, совершенствования продуктов, повышения качества обслуживания и лояльности клиентов.

**Список литературы**

Viktor Mayer-Schönberger, Kenneth Cukier, 2014. Big Data: A Revolution That Will Transform How We Live, Work, and Think

[Rob Thomas](https://www.wiley.com/en-us/search?pq=|relevance|author%3ARob+Thomas), [Patrick McSharry](https://www.wiley.com/en-us/search?pq=|relevance|author%3APatrick+McSharry), 2015. Big Data Revolution: What farmers, doctors and insurance agents teach us about discovering big data patterns

Евгений Шуремов, 2019. Искусственный интеллект и Большие Данные. Без хайпа и наукообразия