О МЕТОДАХ И ИНСТРУМЕНТАХ АНАЛИЗА БОЛЬШИХ ДАННЫХ

ABOUT METHODS AND TOOLS FOR BIG DATA ANALYSIS

Кучкаров Тахир Сафарович, доктор экономических наук, профессор Ташкентского государственного экономического университета

Аннотация. *В статье рассматриваются вопросы использования технологии Big data для обработки, хранения и использования больших данных. Также рассмотрены методы и инструменты искусственного интеллекта и машинного обучения. Наряду с этим перечислены методы обработки структурированной и неструктурированной информации, а также инструментарий больших данных. Представлено современное состояние и тенденции развития технологий Big Data в предприятиях.*

Ключевые слова. *Искусственный интеллект, машинное обучение, Big data, 3V модель данных, нейронные сети.*

Annotation. *In this article the Big data technologies and the process of storing and using big data with AI and ML are discussed. Methods and tools of artificial intelligence and machine learning are also considered. Along with this, the methods for processing structured and unstructured information, as well as big data tools, are listed. The current state and development trends of Big Data technologies in enterprises are presented.*

Keywords. *Artificial intelligence, machine learning, Big data, 3V data model, neural networks*

Введение

С интенсивным развитием Интернета и цифровых технологий во всем мире все более возникают вопросы обработки накапливаемых электронных данных. В современном этапе поиск и совершенствование методов и средств эффективной обработки накопившихся данных в финансовых и бизнес структурах становятся все более актуальными.

Необходимо отметить что, термин Big Data появился в 2008 году. Впервые его употребил редактор журнала Nature - Клиффорд Линч. Он рассказывал про взрывной рост объемов мировой информации и отмечал, что освоить их помогут новые инструменты и более развитые технологии. С увеличением генерации и накопления объема данных остро возникло вопрос их обработки и систематизации.

Big Data – это концепция обработки, анализа и интерпретации огромных массивов данных, которые невозможно обработать с помощью

традиционных методов. Наприме в банковской и финансовой сфере оно представляют собой информацию о клиентах, транзакциях, кредитных историях, а также о киберинцидентах и т.д..

Big data - это совокупность структурированных, неструктурированных и полуструктурированных данных, генерируемые из различных источников данных, такие как информационные системы, ресурсы и порталы, поисковые системы (Google, Yandex, Yahoo и др.), веб сайты, социальные сети (Facebook, Twitter, LinkedIn и др.), сенсорные датчики и автоматизированные системы управления производством (IoT, AI и др.) и другие.(1)

Большие данные (Big Data) стали неотъемлемой частью банковской, финансовой и социальной сферы, поскольку позволяют им принимать более обоснованные решения, основанные на анализе данных. Они помогают предсказывать поведение клиентов, выявлять мошеннические действия, а также снижать операционные издержки и повышать эффективность их деятельности.

Применение современных технологий анализа данных позволяет им привлечь и удерживать клиентов, увеличить прибыль и обеспечить безопасность операций.

Для решения подобных задач, анализа больших данных (Big Data) на сегодняшней широко стали применять различные инструменты и методы, в частности искусственный интеллект (*на англ. Artificial Intelligence, сокр. AI*) и машинное обучение (*на англ. Machine Learning, сокр. ML*). Они стали главными инструментами в обработке больших потоков данных.

В качестве определяющих характеристик для больших данных традиционно выделяют «3V»: объём (Volume), скорость (Velocity), многообразие (Variety). Набор признаков 3V изначально выработан Meta Group в 2001 году вне контекста представлений о больших данных как об определённой серии информационно-технологических методов и инструментов, в нём, в связи с ростом популярности концепции центрального хранилища данных для организаций, отмечалась равнозначимость проблематик управления данными по всем трём аспектам. В дальнейшем появились интерпретации с «четырьмя V» (добавлялась veracity — достоверность, использовалась в рекламных материалах IBM), «пятью V» (в этом варианте прибавляли viability — жизнеспособность, и value — ценность), и даже «семью V» (кроме всего, добавляли также variability — переменчивость, и visualization) (2).

Учитывая разнородность, сложность и объём скапливаемой информации возникает необходимость правильной обработки и

упорядочивание данных для того чтобы использовать их для принятия тех или иных решений. Существующие методы и инструменты сбора, анализа и обработки больших данных можно структурировать данные для из последующей обработки (3,4,5).

Методы исследования

Ниже рассматривается ряд инструментов и методов сбора и анализа структурированных и неструктурированных данных. В зависимости от области применения методы и инструменты различаются и имеют место предназначения.

Изначально в совокупность подходов и технологий Big Data включались средства массово-параллельной обработки неопределённо структурированных данных, такие как СУБД NoSQL, алгоритмы MapReduce и средства проекта Hadoop. В дальнейшем к технологиям больших данных стали относить и другие решения, обеспечивающие сходные по характеристикам возможности по обработке сверхбольших массивов данных, а также некоторые аппаратные средства.

MapReduce — модель распределённых параллельных вычислений в компьютерных кластерах, представленная компанией Google. Согласно этой модели, приложение разделяется на большое количество одинаковых элементарных заданий, выполняемых на узлах кластера и затем естественным образом сводимых в конечный результат.

NoSQL (от англ. Not Only SQL, не только SQL) — общий термин для различных нереляционных баз данных и хранилищ, не обозначает какую-либо одну конкретную технологию или продукт. Обычные реляционные базы данных хорошо подходят для достаточно быстрых и однотипных запросов, а на сложных и гибко построенных запросах, характерных для больших данных, нагрузка превышает разумные пределы и использование СУБД становится неэффективным.

Hadoop — свободно распространяемый набор утилит, библиотек и фреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Считается одной из основополагающих технологий больших данных.

R — язык программирования для статистической обработки данных и работы с графикой. Широко используется для анализа данных и фактически стал стандартом для статистических программ.

Обсуждение

Аппаратные решения. Корпорации Teradata, EMC и др. предлагают аппаратно-программные комплексы, предназначенные для обработки

больших данных. Эти комплексы поставляются как готовые к установке телекоммуникационные шкафы, содержащие кластер серверов и управляющее программное обеспечение для массово-параллельной обработки. Сюда также иногда относят аппаратные решения для аналитической обработки в оперативной памяти, в частности, аппаратно-программные комплексы Hana компании SAP и комплекс Exalytics компании Oracle, несмотря на то, что такая обработка изначально не является массово-параллельной, а объёмы оперативной памяти одного узла ограничиваются несколькими терабайтами.

Справочно: Консалтинговая компания McKinsey, кроме рассматриваемых большинством аналитиков технологий NoSQL, MapReduce, Hadoop, R, включает в контекст применимости для обработки больших данных также технологии Business Intelligence и реляционные системы управления базами данных с поддержкой языка SQL, а также выделяет 11 методов и техник анализа, применимых к большим данным.

Методы класса Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных) — совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных знаний, необходимых для принятия решений. К таким методам, в частности, относятся обучение ассоциативным правилам (association rule learning), классификация (разбиение на категории), кластерный анализ, регрессионный анализ, обнаружение и анализ отклонений и др.

Краудсорсинг — классификация и обогащение данных силами широкого, неопределённого круга лиц, выполняющих эту работу без вступления в трудовые отношения

Смешение и интеграция данных (data fusion and integration) — набор техник, позволяющих интегрировать разнородные данные из разнообразных источников с целью проведения глубинного анализа (например, цифровая обработка сигналов, обработка естественного языка, включая тональный анализ, и др.)

Машинное обучение, включая обучение с учителем и без учителя — использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей

Искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы (genetic algorithm — эвристические алгоритмы поиска, используемые для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации

искомых параметров с использованием механизмов, аналогичных естественному отбору в природе)

Имитационное моделирование (simulation) — метод, позволяющий строить модели, описывающие процессы так, как они проходили бы в действительности. Имитационное моделирование можно рассматривать как разновидность экспериментальных испытаний

Пространственный анализ (spatial analysis) — класс методов, использующих топологическую, геометрическую и географическую информацию, извлекаемую из данных

Статистический анализ — анализ временных рядов, A/B-тестирование (A/B testing, split testing — метод маркетингового исследования; при его использовании контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того чтобы выяснить, какие из изменений улучшают целевой показатель)

Визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа. Очень важный этап анализа больших данных, позволяющий представить самые важные результаты анализа в наиболее удобном для восприятия виде.

Заключение

Применение Big Data в экономике предоставляет компаниям уникальные возможности для оптимизации процессов, принятия обоснованных решений и улучшения клиентского опыта. Несмотря на определенные вызовы и препятствия, оно остается важным инструментом для современных компаний.

Таким образом, вышеуказанные технологии подразумевает работу с информацией колоссального объема и разнообразного состава, часто обновляемой и находящейся в различных источниках в целях увеличения эффективности работы, формирования новых сервисов, создания инновационных маркетинговых инструментов, продвижения продуктов и услуг, оптимизации расходов, улучшения точности прогнозирования и минимизации рисков, и, наконец, повышения конкурентоспособности бизнеса. Главное во всем этом правильно научится использовать инструменты и методы сбора и обработки больших данных в зависимости от области их применения.

Литература

[1]. Deep Transfer Learning for Intelligent Cellular Traffic Prediction Based on Cross-Domain Big Data, 10.1109/JSAC.2019.2904363, IEEE Journal on Selected Areas in Communications

[2]. The 3 V's of Big Data Analytics, 01/04/2019, Stevenn Hansen

[3]. Scientific Data Management in the Age of Big Data: An Approach Supporting a Resilience Index Development Effort, Linda C. Harwell, Stephen F. Hafner

[4]. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, May 2011 | Report

[5] Методы и алгоритмы интеллектуальной системы поддержки принятия решений трейдеров финансовых рынков. // Управление в социальных и экономических системах // Диссертация на соискания кандидата технических наук // Ижевск, 2018 год.