university
Тюменский
индустриальный
университет

ВВЕДЕНИЕ

www.tyuiu.ru

В современном мире данные генерируются и потребляются невиданными ранее темпами.

Откуда берутся эти объемы?

Огромное количество информации генерирует каждый из нас каждый день. Мы разговариваем по телефону, пишем сообщения, ведем блоги, что-то покупаем, что-то фотографируем, что-то отсылаем друзьям, что-то получаем в ответ и т.д. и т.п.

Все это оставляет свой след в информационном пространстве.

Все это где-то хранится и как-то обрабатывается.

Согласно исследованиям, к 2007 году человечество имело возможность хранения информации объемом $2{,}9 \cdot 10^{20}$ байт.

Аналитики компании International Data Corporation (IDC) подсчитали, что в 2020 году общий объём сгенерированных данных составил 64,2 зеттабайта (1 зеттабайт — это миллиард терабайтов).

При этом сообщается, что для дальнейшего использования было сохранено менее 2% информации. Основная часть данных имела временный характер.

Большой объем данных порождают научные эксперименты

Так, в апреле 2016 года в открытый доступ поступили 300 Тбайт экспериментальных данных, полученных на большом адронном коллайдере.

Функционирование многих технических систем также сопровождается сбором большого количества данных. Например, самолет Боинг-787 генерирует около 500 Гбайт данных за один полет.

В настоящее время наиболее быстро растет сегмент данных устройств Интернета вещей. Затем следуют социальные сервисы.

При этом

соотношение уникальных данных (созданных и полученных) к реплицированным данным (скопированным и использованным) составляет примерно 1:9. И наблюдается дальнейшее смещение в сторону реплицированных данных: по прогнозам, к 2024 году это соотношение будет 1:10.

То есть, рост мирового объема данных в большей степени обусловлен данными, которые мы потребляем и анализируем, чем теми, что мы создаем!

В определенный момент информации становится слишком много, и извлекать из нее пользу становится слишком сложно.

Для выделения из накопленных данных полезной информации требуется обработка этих данных!

Анализ данных можно определить как **процесс поиска скрытых закономерностей и генерации новых знаний**.

К **основным задачам анализа** данных можно отнести **прогнозирование**, **классификацию**, **поиск схожих черт**, **выдачу рекомендаций**, **выявление отклонений**.

Анализ данных – междисциплинарная область знаний, находящаяся на стыке **математики**, **теории алгоритмов** и **информационных технологий**.

В англоязычных источниках для обозначения сферы анализа данных используется термины **Data Mining** и **Machine Learning** (машинное обучение).

Согласно энциклопедии **Британника** (http://global.britannica.com), **машинное обучение** является дисциплиной направления «**искусственный интеллект**» (Artificial Intelligence), в свою очередь принадлежащего к области **компьютерных наук** (Computer Science).

Необходимость анализа больших объемов накопленных данных привела к созданию специализированных подразделений во многих компаниях.

Некоторые компании, например **Яндекс**, реализуют собственные образовательные проекты в этой области.

Научные исследования

В сфере анализа данных ведутся активные научные исследования.

Анализ публикаций, индексированных в реферативной базе данных SCOPUS, показывает устойчивый рост количества научных работ.

В последнее время особый интерес в сфере анализа данных вызывают такие направления исследований, как «большие данные» (Big Data) и «глубокое обучение» (Deep Learning).

Scopus

WEB OF SCIENCE™

RU science
ИЗДАТЕЛЬСТВО

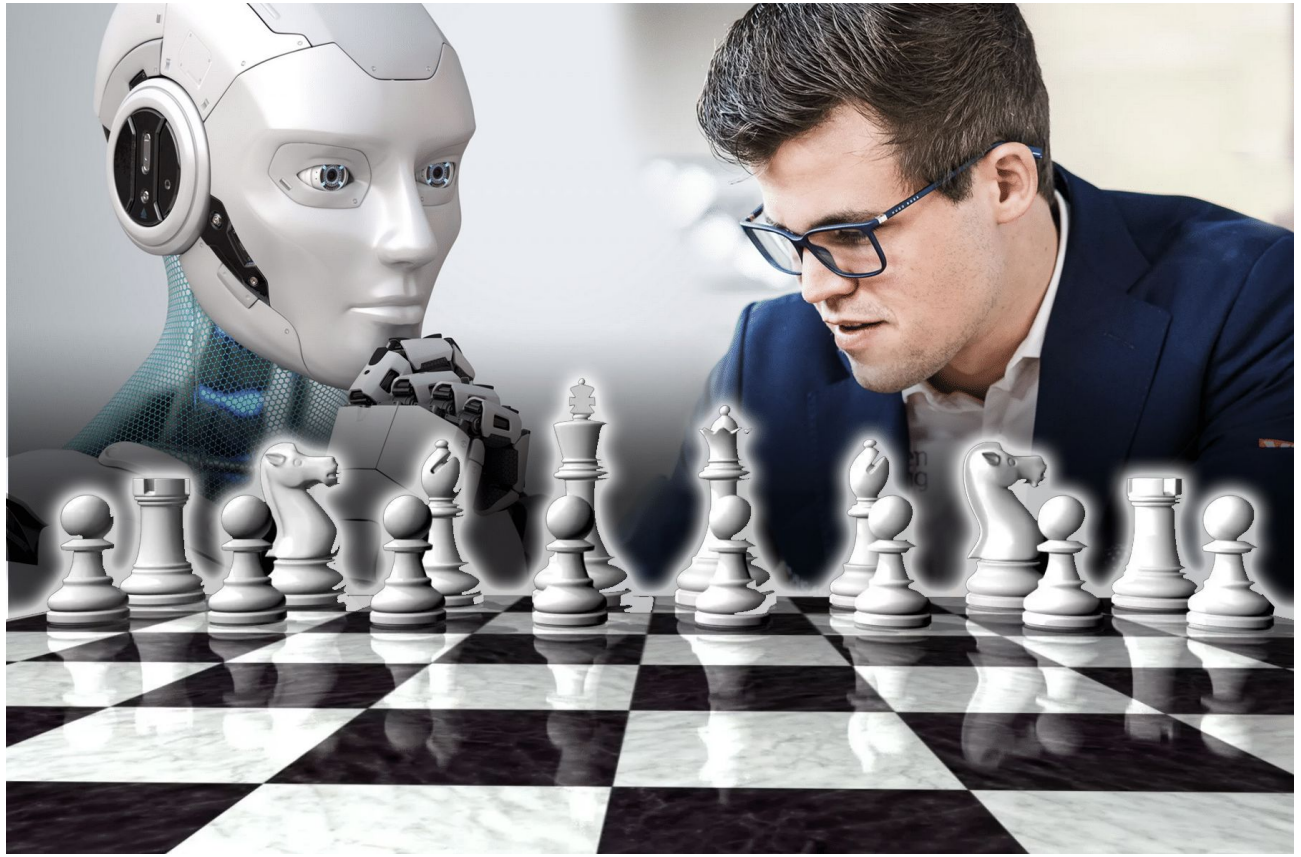
НАУЧНАЯ ЭЛЕКТРОННАЯ БИБЛИОТЕКА
eLIBRARY.RU

Передовые разработки в сфере искусственного интеллекта поражают воображение.

В 1997 году компьютер DeepBlue впервые выиграл матч из шести партий у чемпиона мира по шахматам.

В рамках проекта DeepQA разрабатывается система искусственного интеллекта, позволяющая воспринимать вопросы на естественных языках.

Ведется разработка беспилотных автомобилей, ядром системы управления которых является система искусственного интеллекта. Лидером в этой области можно считать компанию Google с проектом Google Self-Driving Car Project .



Google Self-Driving Car Project является концептом полностью автоматической **машины**, которая управляется при помощи электроники.

Цель **проекта** — не только упростить безопасное передвижение на автомобиле, но и дать шанс людям с ограниченными возможностями самостоятельно использовать **машину** без чужой помощи.

Анализ данных включает три основных этапа

Рис. 2. Этапы анализа данных

Сбор данных – процесс формирования структурированного набора данных в цифровой форме. В некоторых случаях процесс сбора данных может включать также этап оцифровки.

Как правило, оцифрованные данные бывают представлены в виде:

- электронных таблиц в форматах XLS либо ODS;
- текстовых файлов в формате CSV;
- веб-страниц в формате HTML;
- файлов в формате XML;
- базы данных с доступом по технологии JSON либо через специализированный интерфейс (API).

Источники данных

В настоящее время в открытом доступе есть большое количество баз данных, содержащих самые разнообразные сведения. Так, самым большим источником данных по разнообразным показателям стран мира в целом можно считать базу данных Всемирного банка (http://data.worldbank.org), содержащую годовые значения 331 показателя стран мира начиная с 1960 года в форматах HTML, XLS и XML.

Источники данных

Самым большим источником открытых данных по Российской Федерации является «Портал открытых данных Российской Федерации» (http://data.gov.ru), содержащий более 4,1 тыс. наборов данных.

Предполагается, что предоставление свободного доступа к отдельным данным может способствовать повышению качества государственного, регионального и муниципального управления. Принцип открытости получил отдельное название – «открытые данные» (Open Data). В Российской Федерации концепция открытых данных упоминается в Федеральном законе «Об информации, информационных технологиях и о защите информации».

Также большой объем открытых статистических данных содержится в банке данных Федеральной службы государственной статистики (https://www.gks.ru).

Data.gov.ru
ОТКРЫТЫЕ ДАННЫЕ РОССИИ

Федеральная служба
государственной статистики

Программное обеспечение

В основе систем анализа данных лежит программное обеспечение.

При проектировании систем анализа данных могут быть использованы следующие подходы:

- использование «коробочного» программного обеспечения общего назначения (например Microsoft Excel);
- использование программного обеспечения, ориентированного на математические задачи (например Matlab, Octave, R);
- разработка специализированного программного обеспечения с использованием готовых библиотек, включающих наборы специальных функций обработки данных.

При разработке специализированного ПО рекомендуется использовать готовые библиотеки функций обработки данных. Так, для нейросетевого анализа можно применить библиотеку FANN, имеющую версии для языков программирования C#, C++, Java, Python, R, Matlab, а для решения задач обработки изображений – библиотеку OpenCV, имеющую версии для языков Python, Java, Ruby, Matlab и др.

Построение системы анализа данных

Можно предложить следующий общий алгоритм построения системы анализа данных:

1 Постановка задачи.

2 Определение источников данных.

3 Выбор метода и алгоритма обработки данных.

4 Выбор аппаратной платформы.

5 Выбор или разработка программного обеспечения.

6 Верификация построенной системы.

Отметим, что шаги 3 - 5 тесно связаны друг с другом. Например, изменение аппаратной платформы может повлечь необходимость повторной разработки программного обеспечения.

Продолжение следует