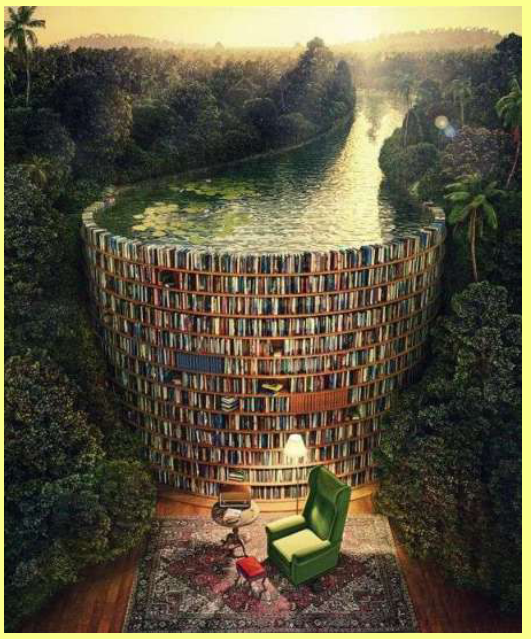
DATA MINING

ИНФОРМАЦИОННЫЙ ПОТОП

По состоянию на начало апреля 2017 года Интернет содержит не менее 4,51 млрд. страниц (те, что проиндексированы поисковыми машинами). Объём всех хранимых данных в 2016 году можно оценить на уровне 2,45 зеттабайта ($2{,}45\times10^{21}$ байтов). Сервис Internet Live Stats наглядно демонстрирует пугающий рост передаваемой по Сети информации: каждую секунду в Google делается более 60 000 поисковых запросов, просматривается 120 000 видео на Youtube, отправляется более 2,5 миллиона электронных писем.



Однако большая часть этих данных не несет какой-либо пользы для конкретного человека (коммерческой компании, государственного или научного учреждения и т. д.) Человек физически не в состоянии освоить и переработать такое количество данных и, к тому же, как правило, требуются не «сырые» данные (raw data), а знания, *информация* (лат. informatio — сведения, разъяснения), содержащаяся в них.

Поэтому для решения задач, связанных с обработкой больших данных (big data), очевидна необходимость применения компьютерных технологий. Сейчас считают, что объем данных в мире удваивается каждые три года, а мощности компьютеров удваиваются каждые полтора года. Несмотря на то, что наши вычислительные мощности пока растут быстрее, чем идёт накопление информации, проблему поиска и, особенно, *получения нужной информации* это в полной мере не решает. Для этого требуется *специальная технология,* объединяющая методы прикладной статистики, распознавания образов, искусственного интеллекта, теории баз данных и др. Её называют

DATA MINING

Data Mining — технология извлечения знаний из больших объемов данных и обнаружения в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны и являются практически полезными. Data Mining можно перевести как «добыча», «раскопка данных» или «добыча полезных ископаемых».

Средствами Data Mining прогнозируют погоду, распознают спам в e-mail, выявляют факты мошенничества, банки решают вопрос о кредитоспособности клиентов, правительства собирают персональную информацию о своих и чужих гражданах (имена, фамилии, адреса, e-mail, номера соцстрахования и т. д. — конечно, исключительно для борьбы с терроризмом и наркотрафиком…), и на основе этой информации осуществляют предсказания возможного поведения людей (подробнее).

Профессия Data Mining сейчас отмечается как одна из самых привлекательных и перспективных в мире, например, по расчётам McKinsey Global Institute к 2018 году в одной только Америке понадобится дополнительно около 190 000 специалистов.

Для отсева большого количества ненужных данных обычно вводится некая функция их полезности. В реальности оценка полезности знания часто имеет субъективный характер, то есть зависит от конкретного пользователя, его целей и предпочтений. Практически процесс Data Mining'а включает следующие этапы:

1. Изучение предметной области
2. Сбор данных.
3. Предварительная обработка данных:
   а. очистка данных;