Data mining - основные понятия и задачи

Лабораторная работа 1

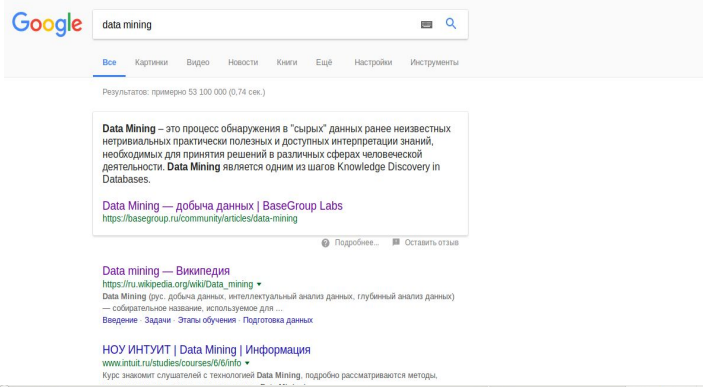
Уровни информации

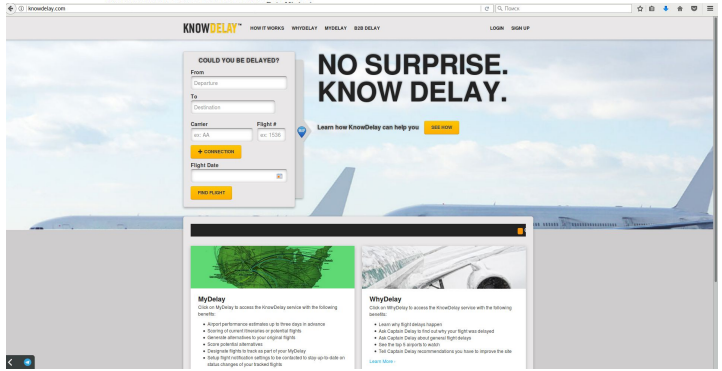
- **исходные данные** – необработанные массивы данных, получаемые в результате наблюдения за некой динамической системой или объектом и отображающие его состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год)
- **информация** – обработанные данные, которые несут в себе некую информационную ценность для пользователя; сырые данные, представленные в более компактном виде (например, результаты поиска)
- **знания** — несут в себе некое ноу-хау, отображают скрытые взаимосвязи между объектами, которые не являются общедоступными (в противном случае, это будет просто информация); данные с большой энтропией (или мерой неопределенности)

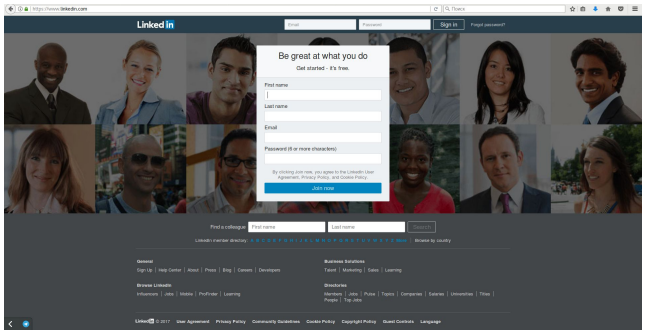
Определения Data Mining

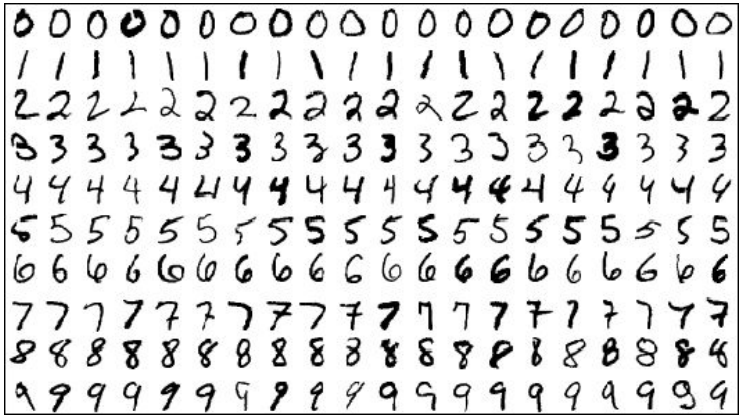
- Извлечение, сбор данных, добыча данных (еще используют Information Retrieval или IR);
- Извлечение знаний, интеллектуальный анализ данных (Knowledge Data Discovery или KDD, Business Intelligence).
- Извлечение знаний из различных источников данных, таких как базы данных, текст, картинки, видео и т.д. Полученные знания должны быть достоверными, полезными и интерпретируемыми.

Применение Data Mining







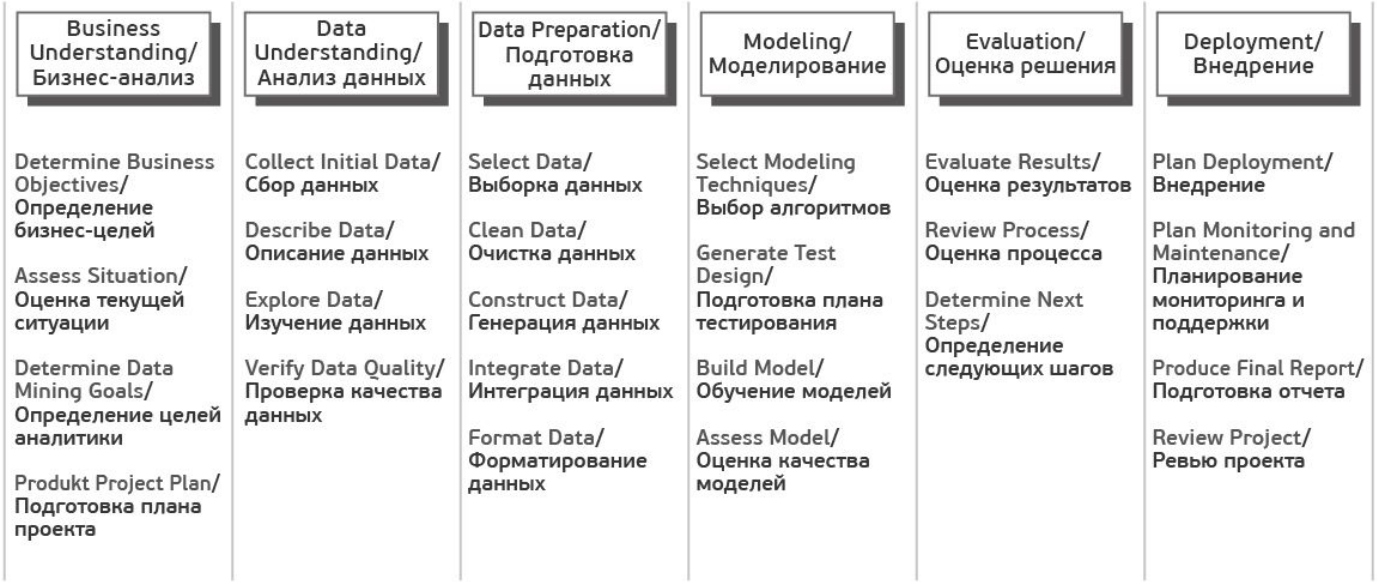


Задачи, решаемые Data Mining

- Классификация — отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов.
- Кластеризация — разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга.
- Сокращение описания — для визуализации данных, упрощения счета и интерпретации, сжатия объемов собираемой и хранимой информации.
- Ассоциация — поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя».
- Прогнозирование – нахождение будущих состояний объекта на основании предыдущих состояний (исторических данных)
- Анализ отклонений — например, выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы.
- Визуализация данных.

CRoss Industry Standard Process for Data Mining (CRISP-DM)

CRoss Industry Standard Process for Data Mining (CRISP-DM)



Программные средства для решения задач Data Mining

RapidMiner
WEKA
R
Orange
KNIME
NLRK
TensorFlow
.
.
.