

Data mining - ОСНОВНЫЕ ПОНЯТИЯ И задачи

Лабораторная работа 1





Уровни информации

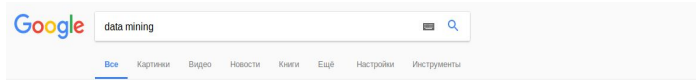
- **исходные данные** – необработанные массивы данных, получаемые в результате наблюдения за некой динамической системой или объектом и отображающие его состояние в конкретные моменты времени (например, данные о котировках акций за прошедший год)
- **информация** – обработанные данные, которые несут 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



Результатов: примерно 53 100 000 (0.74 сек.)

Data Mining – это процесс обнаружения в "сырых" данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. **Data Mining** является одним из шагов Knowledge Discovery in Databases.

[Data Mining — добыча данных | BaseGroup Labs](https://basegroup.ru/community/articles/data-mining)
<https://basegroup.ru/community/articles/data-mining>

[Подробнее...](#) [Оставить отзыв](#)

[Data mining — Википедия](#)

https://ru.wikipedia.org/wiki/Data_mining

Data Mining (рус. добыча данных, интеллектуальный анализ данных, глубинный анализ данных)

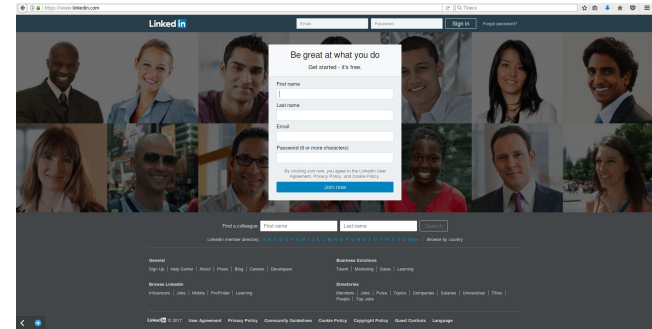
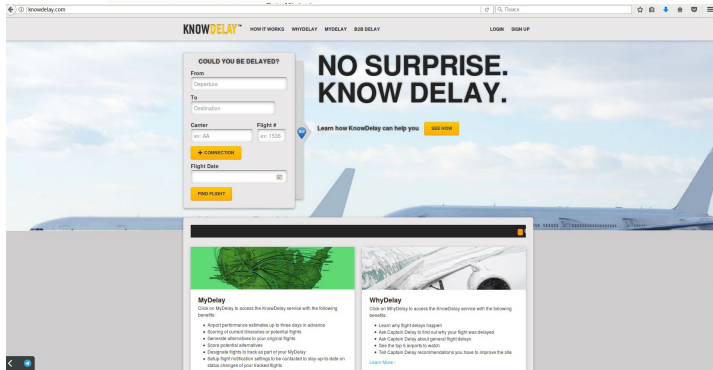
— обработанные название, используемое для ...

Введение · Задачи · Этапы обучения · Подготовка данных

[НОУ ИНТИМ | Data Mining | Информация](#)

www.intim.ru/studies/courses/6/info

Курс знакомит слушателей с технологией Data Mining, подробно рассматриваются методы,





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)

Business Understanding/ Бизнес-анализ	Data Understanding/ Анализ данных	Data Preparation/ Подготовка данных	Modeling/ Моделирование	Evaluation/ Оценка решения	Deployment/ Внедрение
Determine Business Objectives/ Определение бизнес-целей	Collect Initial Data/ Сбор данных	Select Data/ Выборка данных	Select Modeling Techniques/ Выбор алгоритмов	Evaluate Results/ Оценка результатов	Plan Deployment/ Внедрение
Assess Situation/ Оценка текущей ситуации	Describe Data/ Описание данных	Clean Data/ Очистка данных	Generate Test Design/ Подготовка плана тестирования	Review Process/ Оценка процесса	Plan Monitoring and Maintenance/ Планирование мониторинга и поддержки
Determine Data Mining Goals/ Определение целей аналитики	Explore Data/ Изучение данных	Construct Data/ Генерация данных	Build Model/ Обучение моделей	Determine Next Steps/ Определение следующих шагов	Produce Final Report/ Подготовка отчета
Produce Project Plan/ Подготовка плана проекта	Verify Data Quality/ Проверка качества данных	Integrate Data/ Интеграция данных	Assess Model/ Оценка качества моделей		Review Project/ Ревью проекта



Программные средства для решения задач Data Mining

RapidMiner

WEKA

R

Orange

KNIME

NLRK

TensorFlow

.

.

.