Системы интеллектуального анализа данных

Бленда Н.А.
Институт информационных технологий
Челябинский государственный университет
2013г.

Задачи

FALCON (HNC Software, Inc.)

Инструментальное средство для оперативного выявления злоупотреблений с кредитными карточками; более 100 организаций-пользователей отмечают сокращение числа нарушений на 20-30%.

http://www.osp.ru/os/1998/01/179360/

Задачи

Классификатор дебиторских счетов (Internal Revenue Service)

Выявление счетов потенциально платежеспособных дебиторов на основе анализа больших объемов архивных данных по уплате налогов.

http://www.osp.ru/os/1998/01/179360/

Что требуется?

классификация

кластеризация

Выявление фактов, закономерностей

Экспертное мнение

Знания

Данные

Данные

Что является результатом?

Данные

Данные

Знания

Знания

Знания

Знание

результат познания

логическая последовательность суждений и рассматривает знание как основанную на объективной закономерности систему суждений с принципиальной и единой организацией

представляемая в определенной форме информация, ссылаясь на которую делают различные заключения на основании имеющихся данных с помощью логических выводов

Знание

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

http://works.doklad.ru/view/0VYpci5_Juo.html

http://www.osp.ru/os/1998/01/179360/

История Data Mining

- **1960-е гг. –** первая промышленная СУБД система IMS фирмы IBM.
- **1970-е гг. –** Conference on Data System Languages (CODASYL)
- **1980-е гг. –** SQL
- **1990-е гг. –** Data Mining

Data Mining

Data Mining – технология добычи данных

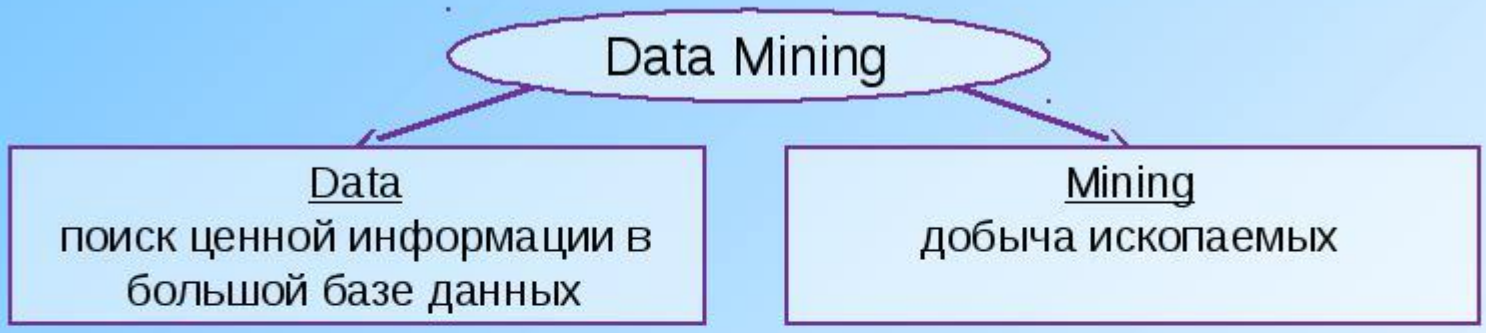
"обнаружение знаний в базах данных" (knowledge discovery in databases)

"интеллектуальный анализ данных"

Специфика современных требований к переработке данных

- Данные имеют неограниченный объем
- Данные являются разнородными (количественными, качественными, текстовыми)
- Результаты должны быть конкретны и понятны
- Инструменты для обработки сырых данных должны быть просты в использовании

Что такое Data Mining?



- Процесс выделения из данных неявной и неструктурированной информации
- Мультидисциплинарная область:
  - прикладная статистика
  - распознавание образов
  - искусственный интеллект
  - теория баз данных


Распознавание образов
Нейросети
Визуализация данных
Статистика
Data Mining
Экспертные системы
Эффективные вычисления
Информацион-ный поиск
Хранилища данных
Теория баз данных
Оперативная аналитическая обработка

Сравним OLAP и Data mining

Примеры формулировок задач при использовании методов OLAP и Data Mining

OLAP	Data Mining
Каковы средние показатели травматизма для курящих и некурящих?	Встречаются ли точные шаблоны в описаниях людей, подверженных повышенному травматизму?
Каковы средние размеры телефонных счетов существующих клиентов в сравнении со счетами бывших клиентов (отказавшихся от услуг телефонной компании)?	Имеются ли характерные портреты клиентов, которые, по всей вероятности, собираются отказаться от услуг телефонной компании?
Какова средняя величина ежедневных покупок по украденной и не украденной кредитной карточке?	Существуют ли стереотипные схемы покупок для случаев мошенничества с кредитными карточками?

Важное положение Data Mining — нетривиальность разыскиваемых шаблонов-найденные шаблоны должны отражать неочевидные, неожиданные (unexpected) регулярности в данных, составляющие так называемые скрытые знания (hidden knowledge).

оперативная аналитическая обработка данных (online analytical processing, OLAP)

Уровни знаний, извлекаемых из данных

SQL

OLAP

Знания и данные

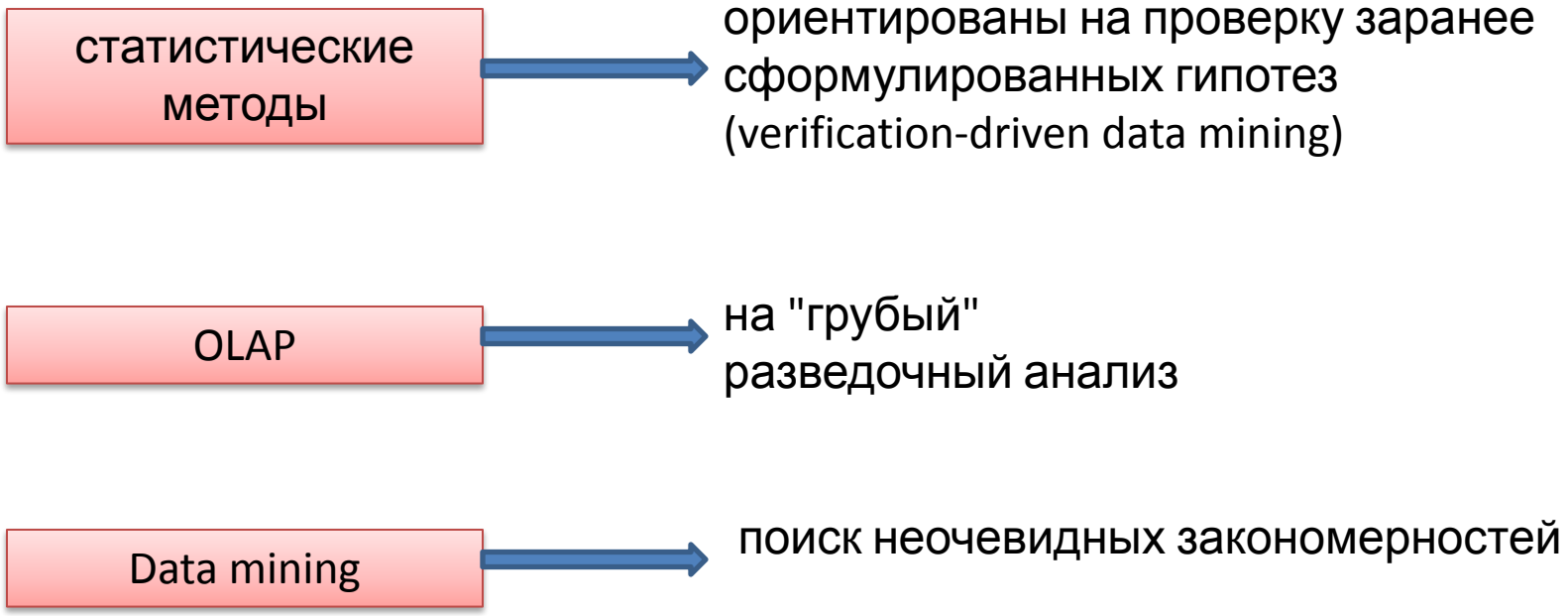
Определение Data mining

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.

Неочевидных - значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем

Определение Data mining

Методы:



Неочевидных - значит, что найденные закономерности не обнаруживаются стандартными методами обработки информации или экспертным путем

Задачи анализа данных

Классификация (Classification)

Кластеризация (Clustering)

Ассоциация (Associations)

Последовательность (Sequence)

Прогнозирование (Forecasting)

Определение отклонений или выбросов (Deviation Detection)

Оценивание (Estimation)

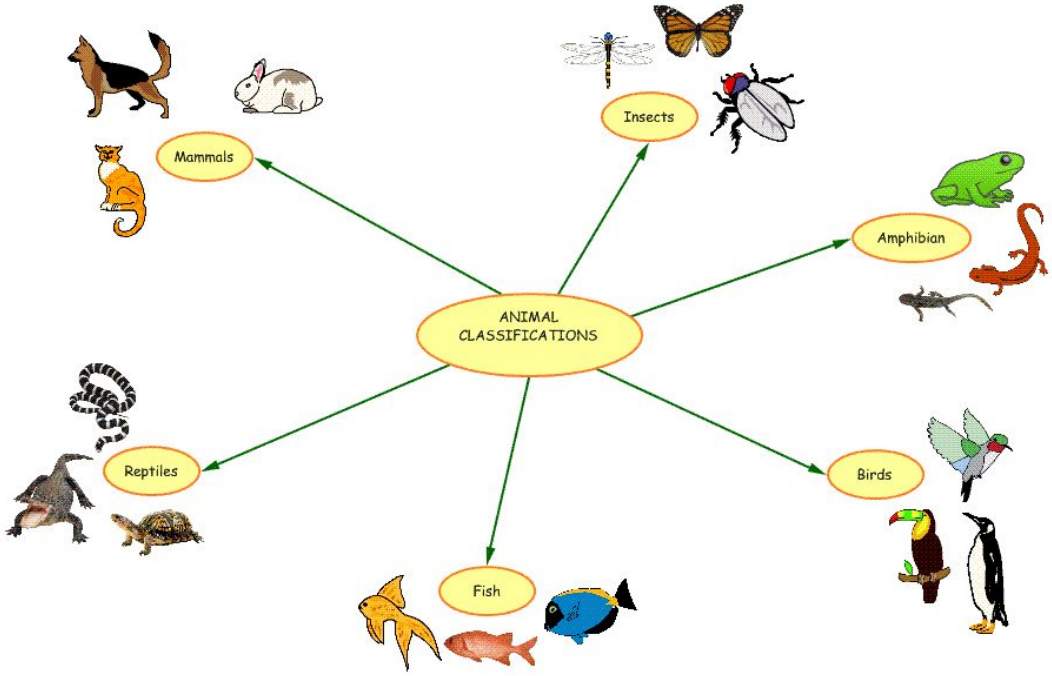
Анализ связей (Link Analysis)

Визуализация (Visualization, Graph Mining)

Подведение итогов (Summarization)

Задачи анализа данных

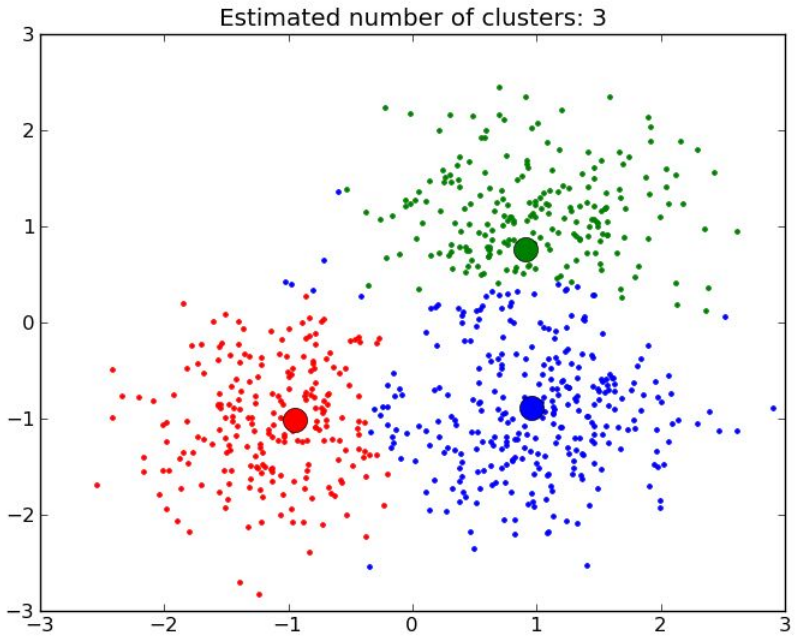
Классификация (Classification)



Для решения задачи классификации могут использоваться методы: ближайшего соседа (Nearest Neighbor); k-ближайшего соседа (k-Nearest Neighbor); байесовские сети (Bayesian Networks); индукция деревьев решений; нейронные сети (neural networks)

Задачи анализа данных

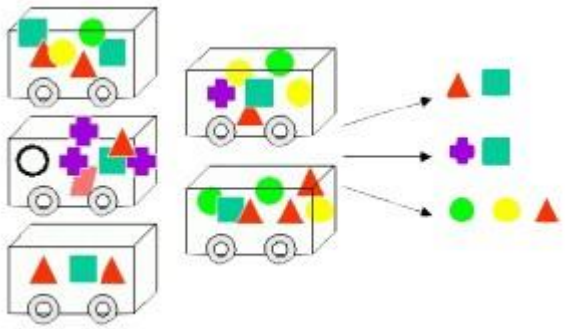
Кластеризация (Clustering)



особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Задачи анализа данных

Ассоциация (Associations)



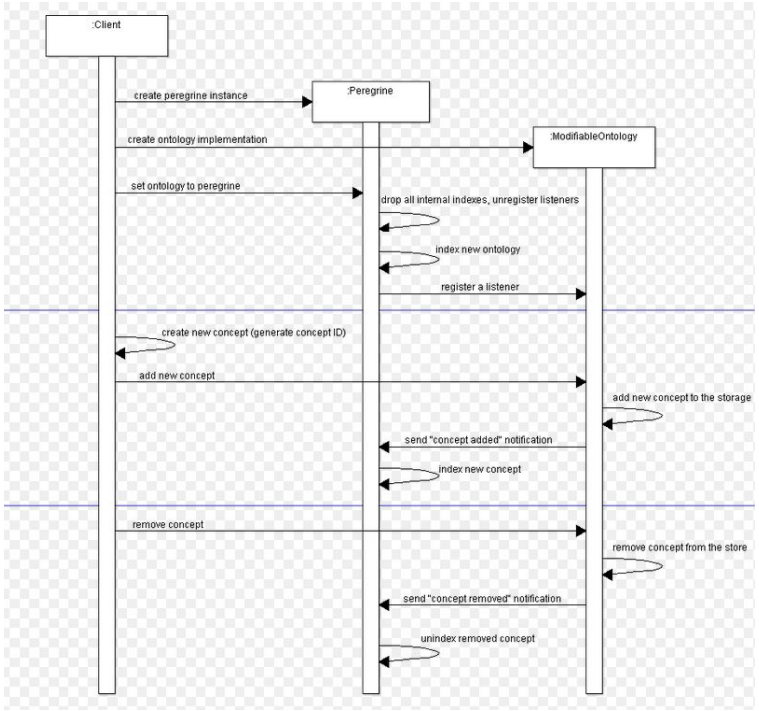
В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных. Отличие ассоциации от двух предыдущих задач Data Mining: поиск закономерностей осуществляется не на основе свойств анализируемого объекта, а между несколькими событиями, которые происходят одновременно. Наиболее известный алгоритм решения задачи поиска ассоциативных правил - алгоритм Apriori.

Задачи анализа данных

Последовательность (Sequence)

последовательная ассоциация (sequential association)
Последовательность позволяет найти временные закономерности между транзакциями.

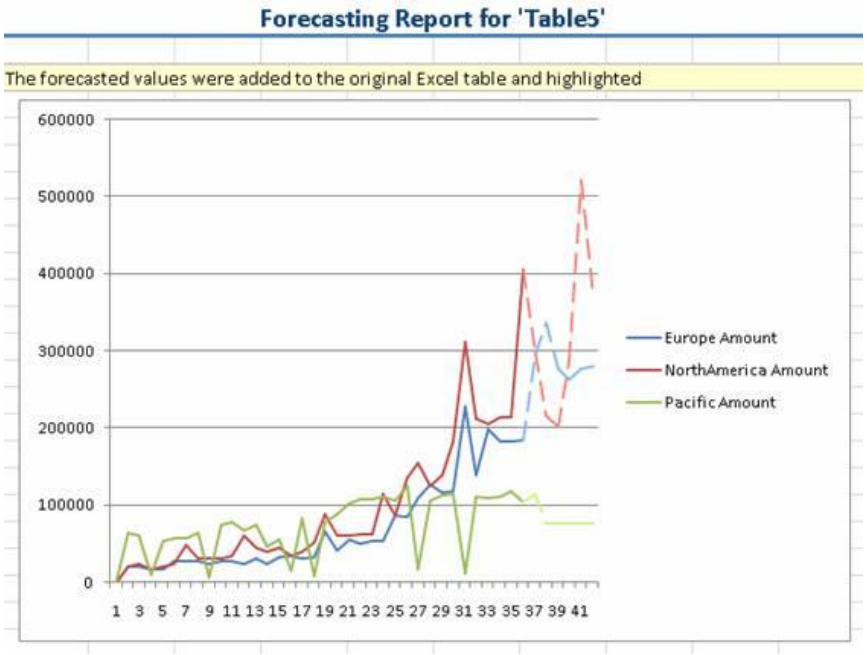
Ассоциация с временными интервалами =0



Пример. После покупки квартиры жильцы в 60% случаев в течение двух недель приобретают холодильник, а в течение двух месяцев в 50% случаев приобретается телевизор.

Задачи анализа данных

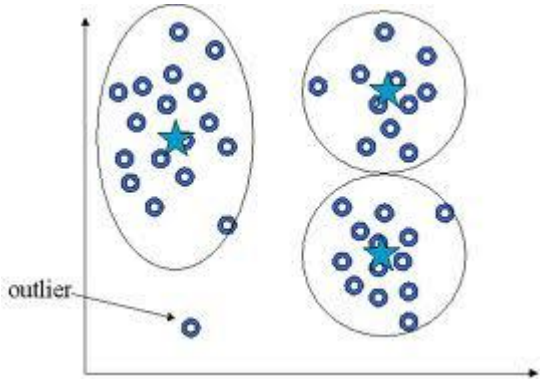
Прогнозирование (Forecasting)

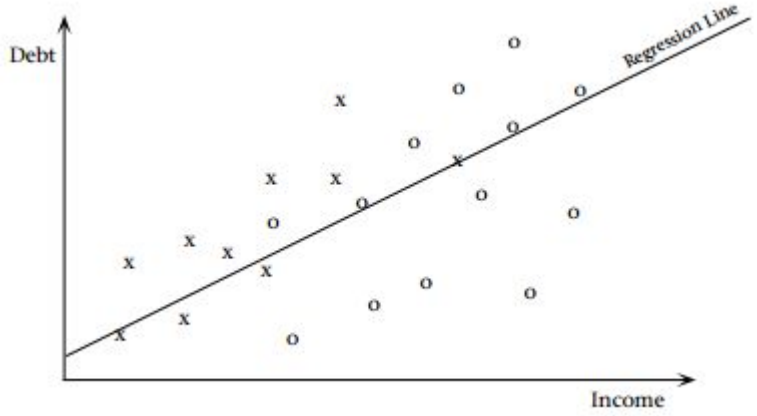


Для решения таких задач широко применяются методы математической статистики, нейронные сети и др.

Задачи анализа данных

Определение отклонений или выбросов (Deviation Detection)





Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

Задачи анализа данных

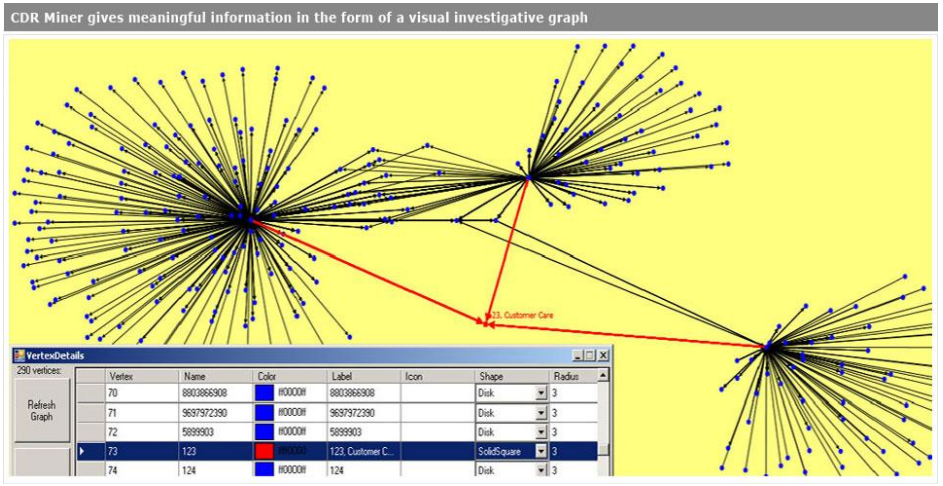
Оценивание (Estimation)

Предположим, что состояние системы в момент времени t определяется, вообще говоря, случайным вектором x(t) k Rn, где t $ t0 и t0 - заданный начальный момент времени. При каждом t $ t0 наблюдается другой случайный вектор, y(t) k Rm. Требуется при каждом t построить такую функцию , зависящую от - результатов измерений y(s), t0 # # s # t, которая в некотором смысле наилучшим образом аппроксимировала бы неизвестный фазовый вектор x(t). При этом функция обычно именуется оценкой вектора x(t).

Задача оценивания сводится к предсказанию непрерывных значений признака

Задачи анализа данных

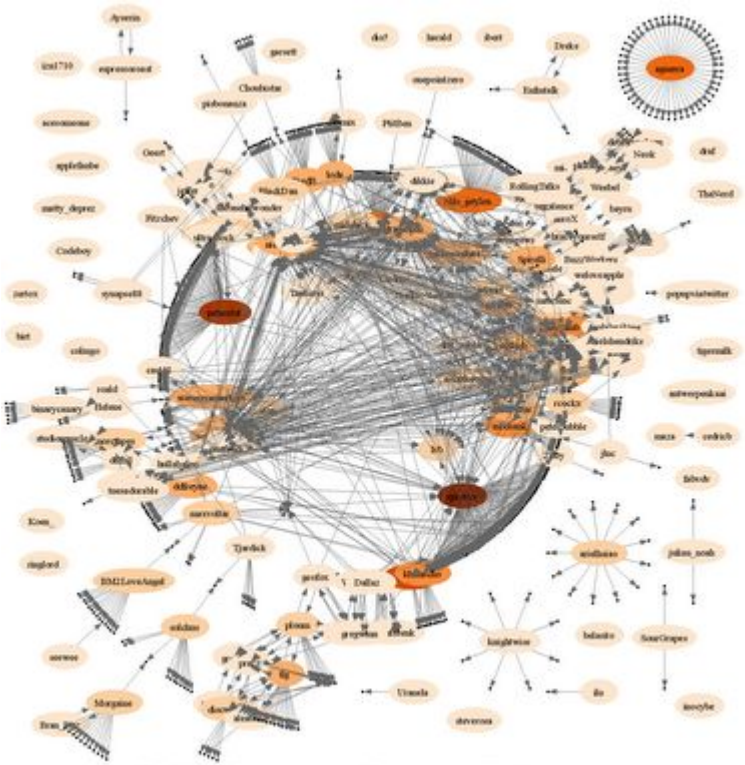
Анализ связей (Link Analysis)



задача нахождения зависимостей в наборе данных.

Задачи анализа данных

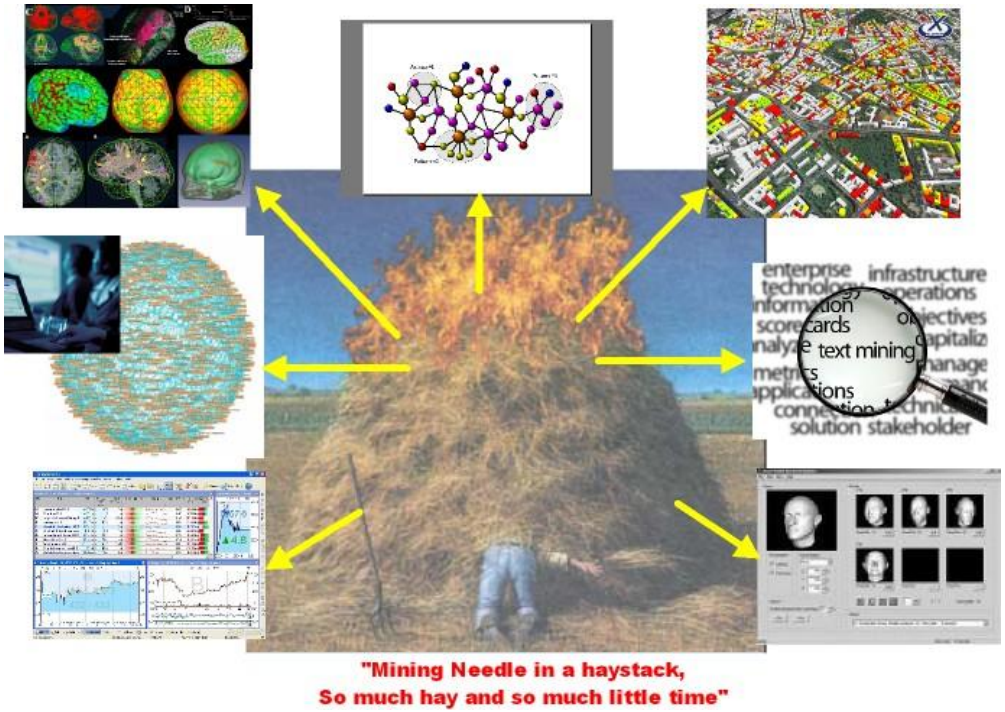
Визуализация (Visualization, Graph Mining)



Twitter Friends van Belgische Twitteraars

Задачи анализа данных

Подведение итогов (Summarization)

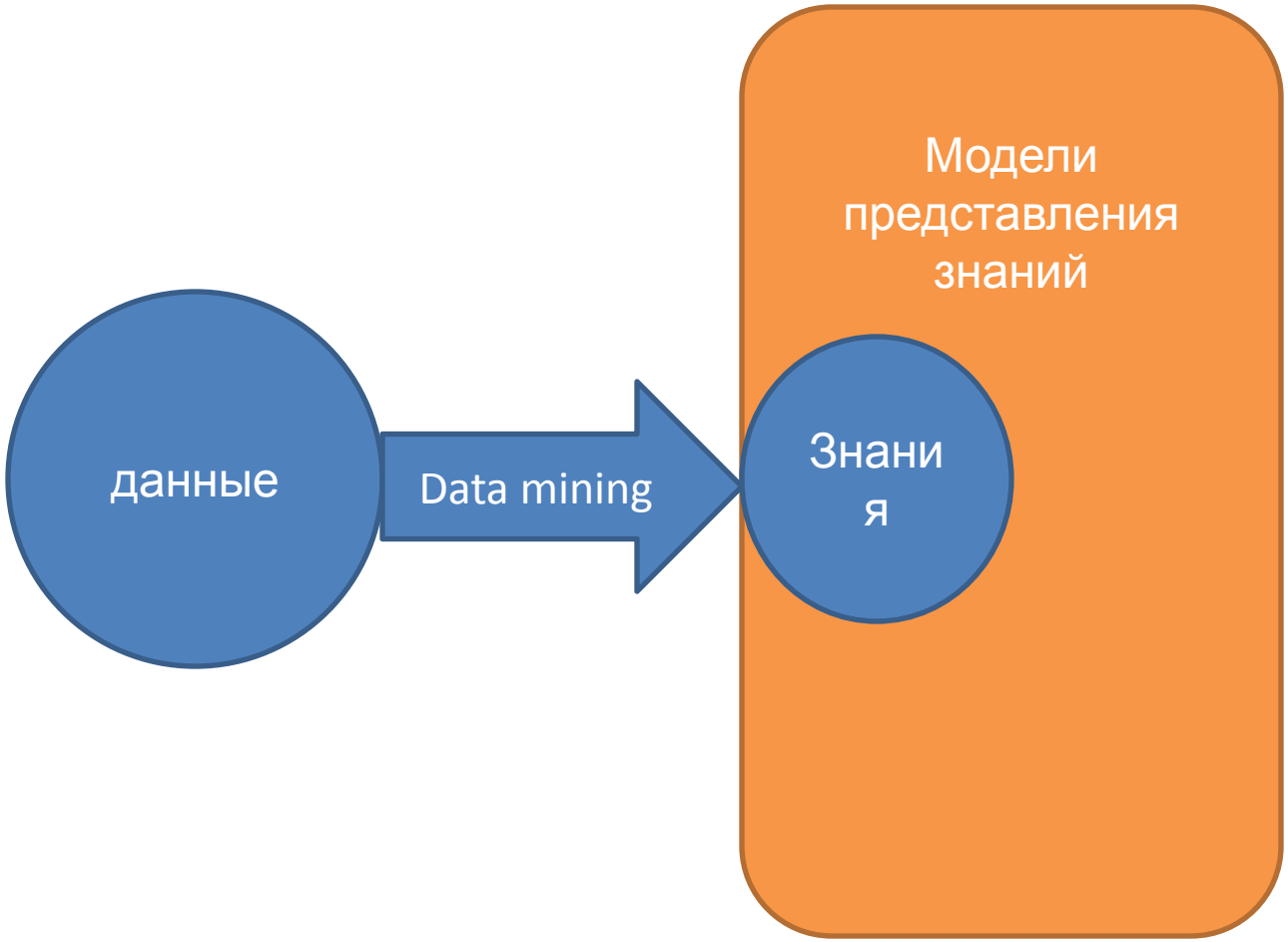


"Mining Needle in a haystack,
So much hay and so much little time"

задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

Закономерности, которые выявляет Data mining

Популярные продукты для Data Mining


данные
Data mining
Знани
я
Модели
представления
знаний


Модели представления знаний
Декларативные
Знания — это данные, так или иначе структурированные. Интерпретация структур и выполнение операций над ними — функция программных средств
Процедурные
Знания представляются в виде структур данных, но при этом с элементами структур ассоциируются некоторые специализированные процедуры
Логические
Вся система знаний, необходимая для решения прикладных задач, рассматривается как совокупность фактов (утверждений). Факты представляются как формулы в некоторой логике
Продукционные
Подобные системы управляются с помощью правил продукций и состоят из трех основных частей: множества правил продукций, базы данных (множества фактов) и интерпретации (означивания)
Аппарат семантических сетей
Проблемная среда рассматривается как совокупность объектов (сущностей) и связей (отношений) между ними
Логика высказываний
Высказывание рассматривается как единое целое, не обладающее внутренней структурой. Истинность или ложность высказывания фиксирована
Исчисление предикатов
Основной объект — переменное высказывание (предикат), истинность или ложность которого зависит от значения входящих в него переменных
Семантические сети
Система знаний, отображается семантической сетью — ориентированным графом, состоящим из поименованных вершин и ребер, или совокупностью таких сетей
Фреймы
Фрейм — фрагмент семантической сети, предназначенный для описания объекта (ситуации) проблемной среды со всей совокупностью присущих ему свойств

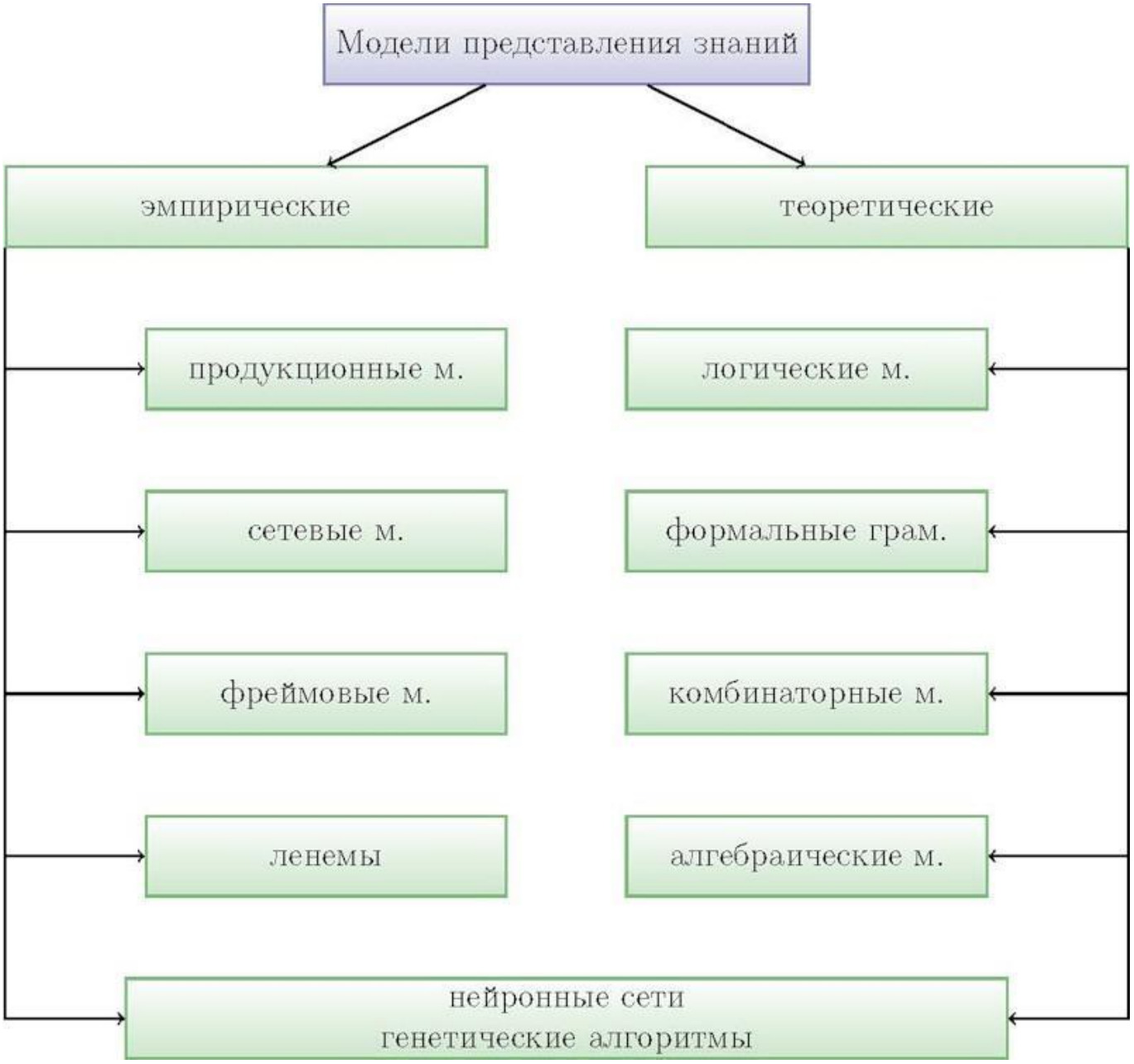

Модели представления знаний
эмпирические
теоретические
продукционные м.
логические м.
сетевые м.
формальные грам.
фреймовые м.
комбинаторные м.
ленемы
алгебраические м.
нейронные сети
генетические алгоритмы

Рис. 2.3. Классификация методов представления знаний