Data Mining

1. Составитель
2. Введение в Data Mining
3. Методы
4. Области применения

Составитель

- Алёшин Владислав, ИТ-7, 1 курс

Возникновение Data Mining. Способствующие факторы

- совершенствование аппаратного и программного обеспечения;
- совершенствование технологий хранения и записи данных;
- накопление большого количества ретроспективных данных;
- совершенствование алгоритмов обработки информации.

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

Понятие 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.*

Gregory Piatetsky-Shapiro

- Это технология, которая предназначена для поиска в больших объемах данных неочевидных, объективных и полезных на практике закономерностей.

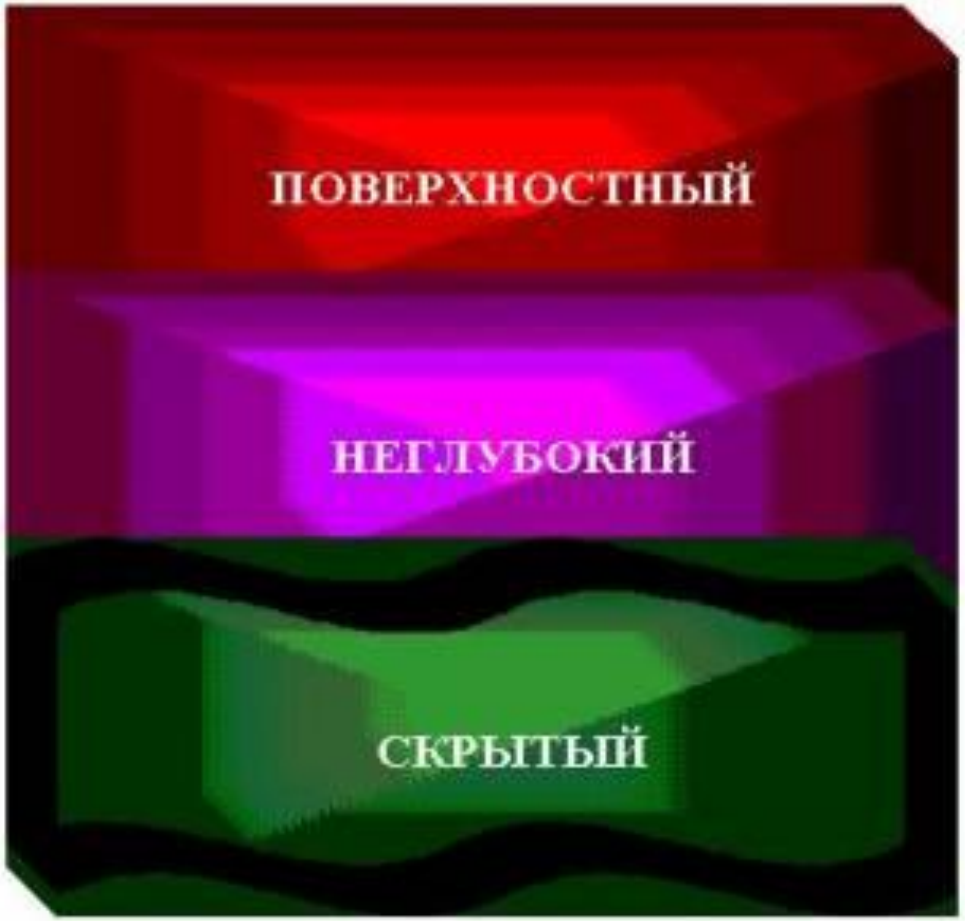
Мультидисциплинарность

Составные части Data Mining

Уровни знаний, извлекаемых из данных

Технологии «сверху-вниз»

Технологии «снизу-вверх»



Аналитические инструменты

Язык простых запросов

Оперативная аналитическая обработка

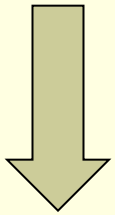
Data Mining «Раскопка данных»

Задачи Data Mining

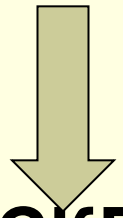
- Классификация
- Кластеризация
- Прогнозирование
- Ассоциация
- Визуализация
- анализ и обнаружение отклонений
- Оценивание
- Анализ связей
- Подведение итогов

Стадии Data Mining

СВОБОДНЫЙ ПОИСК (в том числе ВАЛИДАЦИЯ)



ПРОГНОСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ



АНАЛИЗ ИСКЛЮЧЕНИЙ

Методы Data Mining. Технологические методы.

- Непосредственное использование данных, или сохранение данных:
  кластерный анализ, метод ближайшего соседа, метод k-ближайшего соседа, рассуждение по аналогии
- Выявление и использование формализованных закономерностей, или дистилляция шаблонов:
  логические методы; методы визуализации; методы кросс-табуляции; методы, основанные на уравнениях

Методы Data Mining. Статистические методы.

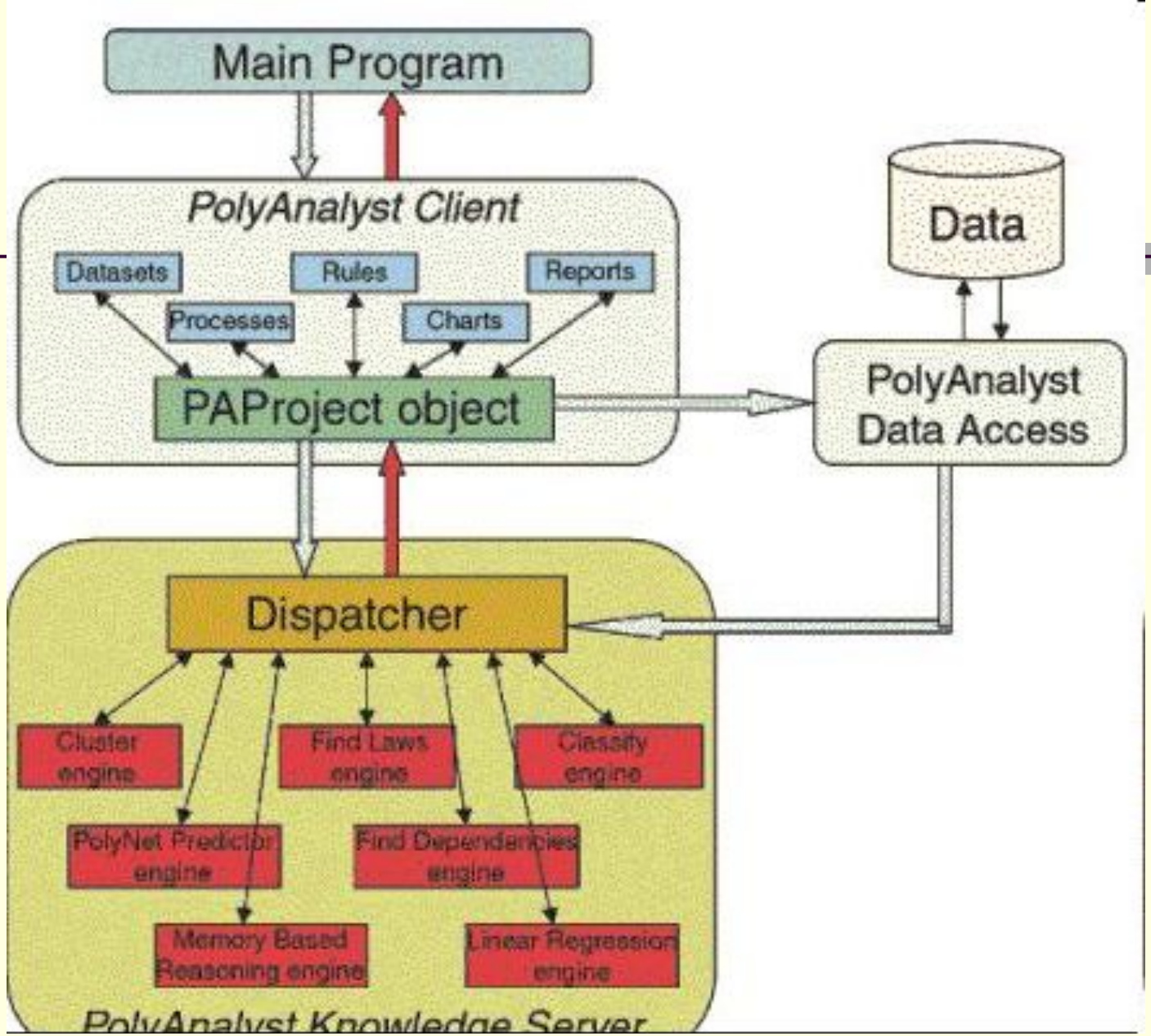
- Дескриптивный анализ и описание исходных данных.
- Анализ связей (корреляционный и регрессионный анализ, факторный анализ, дисперсионный анализ).
- Многомерный статистический анализ (компонентный анализ, дискриминантный анализ, многомерный регрессионный анализ, канонические корреляции и др.).
- Анализ временных рядов (динамические модели и прогнозирование).

Методы Data Mining. Кибернетические методы.

- Искусственные нейронные сети (распознавание, кластеризация, прогноз);
- Эволюционное программирование (в т.ч. алгоритмы метода группового учета аргументов);
- Генетические алгоритмы (оптимизация);
- Ассоциативная память (поиск аналогов, прототипов);
- Нечеткая логика;
- Деревья решений;
- Системы обработки экспертных знаний.

Визуализация инструментов Data Mining.

- Для деревьев решений - визуализатор дерева решений, список правил, таблица сопряженности.
- Для нейронных сетей - в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.
- Для карт Кохонена: карты входов, выходов, другие специфические карты.
- Для линейной регрессии - линия регрессии.
- Для кластеризации: дендрограммы, диаграммы рассеивания.



Архитектура системы PolyAnalyst

Проблемы и вопросы

- Data Mining не может заменить аналитика!
- Сложность разработки и эксплуатации приложения Data Mining. Основные аспекты:
  - Квалификация пользователя
  - Сложность подготовки данных
  - Большой процент ложных, недостоверных или бессмысленных результатов
  - Высокая стоимость
  - Наличие достаточного количества репрезентативных данных

Области применения Data mining

- ***Database marketers*** - Рыночная сегментация, идентификация целевых групп, построение профиля клиента
- ***Банковское дело*** - Анализ кредитных рисков, привлечение и удержание клиентов, управление ресурсами
- ***Кредитные компании*** - Детекция подлогов, формирование "типичного поведения" обладателя кредитки, анализ достоверности клиентских счетов , cross-selling программы
- ***Страховые компании*** - Привлечение и удержание клиентов, прогнозирование фингансовых показателей
- ***Розничная торговля*** - Анализ деятельности торговых точек, построение профиля покупателя, управление ресурсами
- ***Биржевые трейдеры*** - Выработка оптимальной торговой стратегии, контроль рисков

Области применения Data mining. Продолжение.

- ***Телекоммуникация и энергетика*** - Привлечение клиентов, ценовая политика, анализ отказов, предсказание пиковых нагрузок, прогнозирование поступления средств
- ***Налоговые службы и аудиторы*** - Детекция подлогов, прогнозирование поступлений в бюджет
- ***Фармацевтические компании*** - Предсказание результатов будущего тестирования препаратов, программы испытания
- ***Медицина*** - Диагностика, выбор лечебных воздействий, прогнозирование исхода хирургического вмешательства
- ***Управление производством*** - Контроль качества, материально-техническое обеспечение, оптимизация технологического процесса
- ***Ученые и инженеры*** - Построение эмпирических моделей, основанных на анализе данных, решение научно-технических задач

Перспективы технологии Data Mining.

- выделение типов предметных областей с соответствующими им эвристиками
- создание формальных языков и логических средств, с помощью которых будет формализованы рассуждения
- создание методов Data Mining, способных не только извлекать из данных закономерности, но и формировать некие теории, опирающиеся на эмпирические данные;
- преодоление существенного отставания возможностей инструментальных средств Data Mining от теоретических достижений в этой области.

Литература по Data Mining

- "Wikipedia about Data Mining" (http://en.wikipedia.org/wiki/Data_mining)
- "Data Mining Tutorials" (http://www.eruditionhome.com/datamining/tut.html)
- "Thearling intro paper" (http://www.thearling.com/text/dmwhite/dmwhite.htm)
- "Что такое Data mining?“ (http://www.megaputer.ru/doc.php?classroom/whatis_dm/whatis_dm.html)
- "INTUIT.ru: Учебный курс - Data Mining“ (http://www.intuit.ru/department/database/datamining/)
- "Data Mining - подготовка исходных данных“ (http://www.basegroup.ru/tasks/datamining_prepare.htm)