Интеллектуальный анализ данных. АП Deductor (Loginom)

Лекция 6 (1 час)

Емельянова М.Г.

Понятие интеллектуального анализа данных

Три группы направлений в области интеллектуального анализа данных: **Data Mining**, **Machine Learning** и **Knowledge Discovery in Databases** (KDD).

KDD включает Data Mining как ядро и при этом может использовать методы Machine Learning.

KDD – процесс получения из данных знаний в виде зависимостей, правил и моделей позволяющих моделирование и прогнозирование различных процессов.

Термин Data Mining был введен для обозначения совокупности методов автоматизированного решения сложных задач с целью извлечения полезной информации из больших баз данных в виде свойств, группировок и зависимостей в данных и дальнейшего применения этой информации для получения экономической или иной выгоды.

Понятие интеллектуального анализа данных

Совокупность методов, образующих направление интеллектуального анализа данных и место Data Mining среди них.

Сравнение статистики, машинного обучения и Data Mining

Статистика

Более, чем Data Mining, базируется на теории.

Более сосредотачивается на проверке гипотез.

Машинное обучение

Более эвристично.

Концентрируется на улучшении работы агентов обучения.

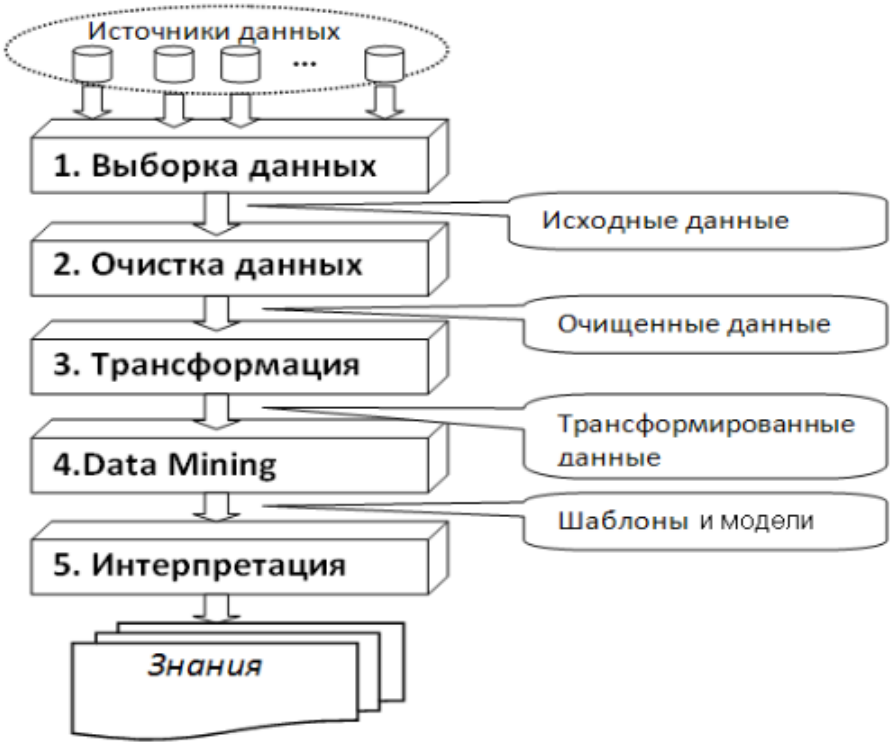
Data Mining

Интеграция теории и эвристик.

Сконцентрирована на едином процессе анализа данных, включает очистку данных, обучение, интеграцию и визуализацию результатов.

Понятие интеллектуального анализа данных

KDD включает в себя **этапы** подготовки данных, предобработки (очистки) данных, трансформации данных, построения моделей (применения методов Data Mining) и интерпретации полученных результатов. Ядром этого процесса являются методы Data Mining, позволяющие обнаруживать закономерности и знания.



Этапы KDD

Первый этап. Выборка данных.

Этот этап заключается в подготовке набора данных, в том числе из различных источников, выбора значимых параметров и т.д. Для этого должны быть различные инструменты доступа к различным источникам данных – конверторы, запросы, фильтрация данных и т.п. В качестве источника рекомендуется использовать специализированное **хранилище данных**, агрегирующее всю необходимую для анализа информацию.

Хранилище данных

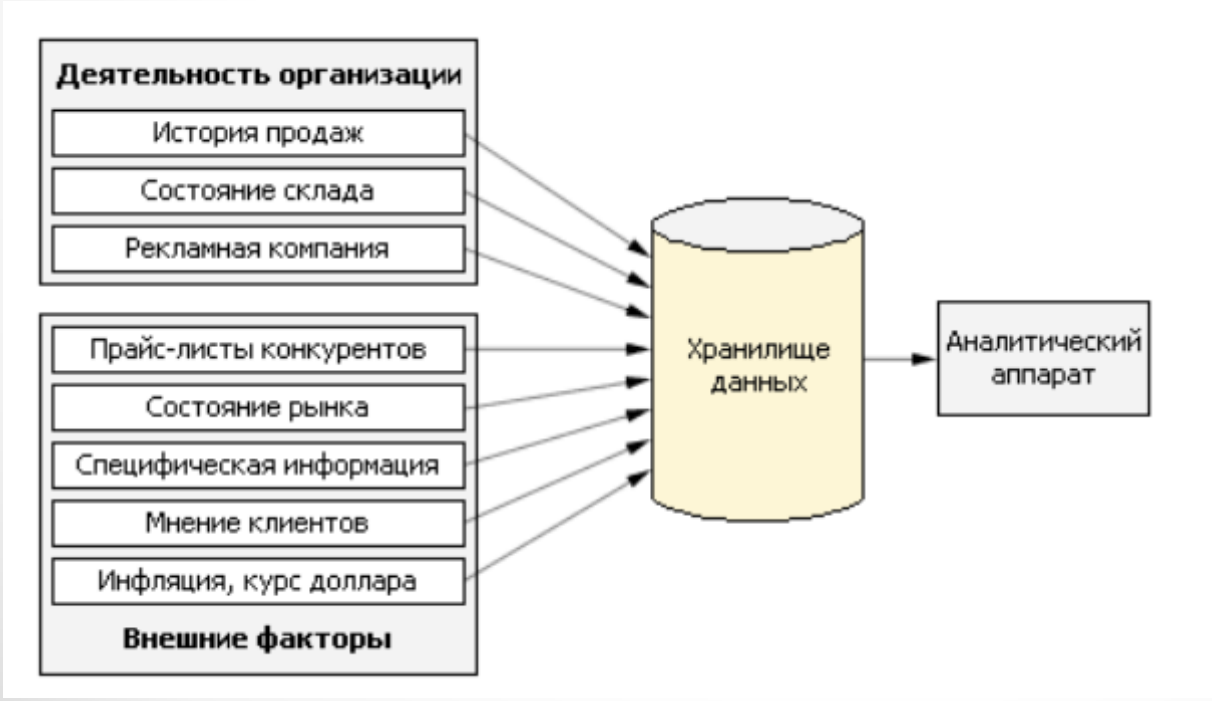
Хранилище данных – разновидность систем хранения, ориентированная на поддержку процесса анализа данных, обеспечивающая целостность, непротиворечивость и хронологию данных, а также высокую скорость выполнения аналитических запросов.

Хранилище данных – это специальным образом систематизированная информация из разнородных источников (базы данных учетных систем компании, маркетинговые данные, мнения клиентов, исследования конкурентов и т.п.), необходимая для обработки с целью принятия стратегически важных решений в деятельности компании.

Хранилище данных

Например, для прогнозирования объемов продаж может потребоваться различная и разнородная информация.

Хранилище данных консолидирует всю необходимую информацию для осуществления задач стратегического управления в среднесрочном и долгосрочном периоде.



Хранилища данных

Архитектуры хранилищ данных:

- многомерные;
- реляционные;
- гибридные;
- виртуальные.

Многомерная модель данных, лежащая в основе построения **многомерных хранилищ данных**, опирается на концепцию многомерных кубов, или гиперкубов. Они представляют собой упорядоченные многомерные массивы, которые называют OLAP-кубами (On-Line Analytical Processing – оперативная аналитическая обработка).

Технология OLAP представляет собой методику оперативного извлечения нужной информации из больших массивов данных и формирования соответствующих отчетов.

Этапы KDD

Второй этап. Очистка данных.

Реальные данные для анализа редко бывают хорошего качества. Поэтому для эффективного применения методов Data Mining следует обратить серьезное внимание на вопросы предобработки данных. Данные могут содержать пропуски, шумы, аномальные значения и т.д. Кроме того, данные могут быть противоречивы, избыточны, недостаточны, содержать ошибки и т.д.

Третий этап. Трансформация данных.

Этот шаг необходим для тех методов, которые требуют, чтобы исходные данные были в каком-то определенном виде. Различные алгоритмы анализа требуют специальным образом подготовленные данные.

Этапы KDD

Четвёртый этап. Data Mining.

На этом этапе строятся модели, в которых применяются различные алгоритмы для нахождения знаний. Это нейронные сети, деревья решений, алгоритмы кластеризации и установления ассоциаций и т.д.

Пятый этап. Интерпретация.

На данном этапе осуществляется применение пользователем полученных моделей (знаний) в бизнес приложениях. Для оценки качества полученной модели нужно использовать как формальные методы, так и знания аналитика.

Понятие Data Mining

Data Mining шаг процесса KDD.

Основателем и одним из идеологов Data Mining считается Пятецкий-Шапиро. Впервые термин был введен в 1989 году на одном из семинаров, посвященных технологиям поиска знаний в базах данных, проводимых в рамках Международной конференции по искусственному интеллекту (International Joint Conference on Artificial Intelligence) IJCAI-89.

Data Mining – обнаружение в «сырых» данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Информация, найденная в процессе применения методов Data Mining, должна быть нетривиальной и ранее неизвестно. Знания должны описывать новые связи между свойствами, предсказывать значения одних признаков на основе других.

Задачи и методы Data Mining

Задачи, решаемые методами Data Mining: классификация, прогнозирования (регрессия), кластеризация, ассоциация и др.

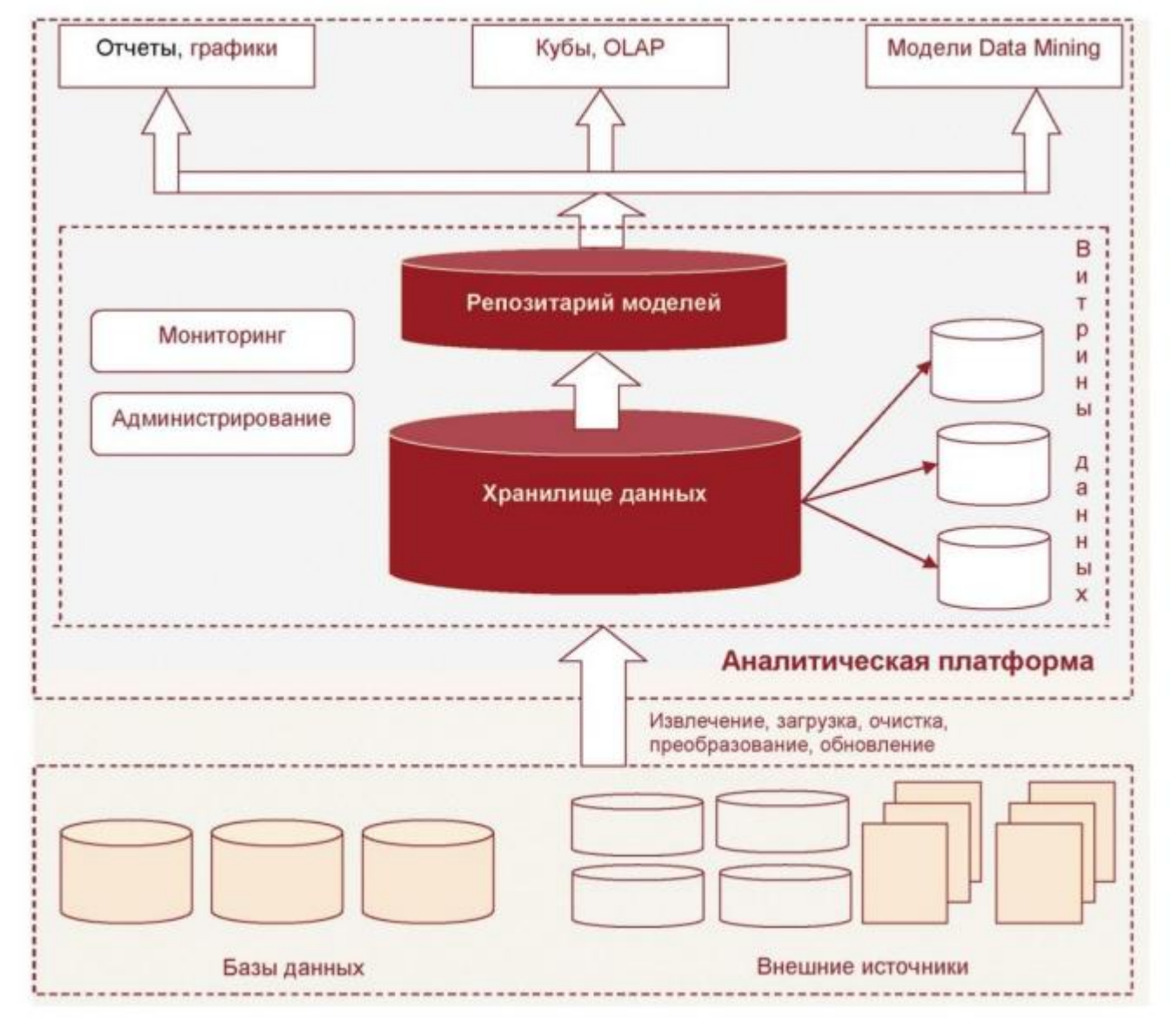
Ввиду того, что Data Mining развивалась и развивается на стыке таких дисциплин, как математическая статистика, теория информации, машинное обучение и базы данных, вполне закономерно, что большинство алгоритмов и методов Data Mining были разработаны на основе различных методов из этих дисциплин.

https://wiki.loginom.ru/articles/data-mining.html

Средства реализации интеллектуального анализа данных

Аналитическая платформа – специализированное программное решение (или набор решений), которое содержит в себе все инструменты для извлечения закономерностей из «сырых» данных: средства консолидации информации в едином источнике (хранилище данных), извлечения, преобразования, трансформации данных, алгоритмы Data Mining, средства визуализации и распространения результатов среди пользователей, а также возможности «конвейерной» обработки новых данных.

Типовая схема системы ИАД на базе аналитической платформы



Аналитическая платформа Deductor

Аналитическая платформа Deductor (Loginom) является одним из программных продуктов, реализующих методы ИАД.

Аналитическая платформа Deductor разработана фирмой BaseGroup Labs (https://basegroup.ru/).

Аналитическая платформа нового поколения – Loginom.

Аналитическая платформа Deductor

Аналитическая платформа Deductor состоит из пяти частей:

- Deductor Warehouse;
- Deductor Studio;
- Deductor Viewer;
- Deductor Server;
- Deductor Client.

Deductor Warehouse – многомерное хранилище данных, аккумулирующее всю необходимую для анализа предметной области информацию. Использование единого хранилища позволяет обеспечить непротиворечивость данных, их централизованное хранение и автоматически обеспечивает всю необходимую поддержку процесса анализа данных.

Аналитическая платформа Deductor

Deductor Studio – программа, реализующая функции импорта, обработки, визуализации и экспорта данных. Deductor Studio может функционировать и без хранилища данных, получая информацию из других источников.

В Deductor Studio включен полный набор механизмов, позволяющий получить информацию из произвольного источника данных, провести весь цикл обработки (очистку, трансформацию данных, построение моделей), отобразить полученные результаты наиболее удобным образом (OLAP, диаграммы, деревья…) и экспортировать результаты на сторону. Это полностью соответствует концепции извлечения знаний из баз данных (KDD).

Аналитическая платформа Deductor

Deductor Viewer – рабочее место конечного пользователя.

Deductor Server – служба, обеспечивающая удаленную аналитическую обработку данных через компьютерную сеть.

Deductor Client – клиент доступа к Deductor Server.

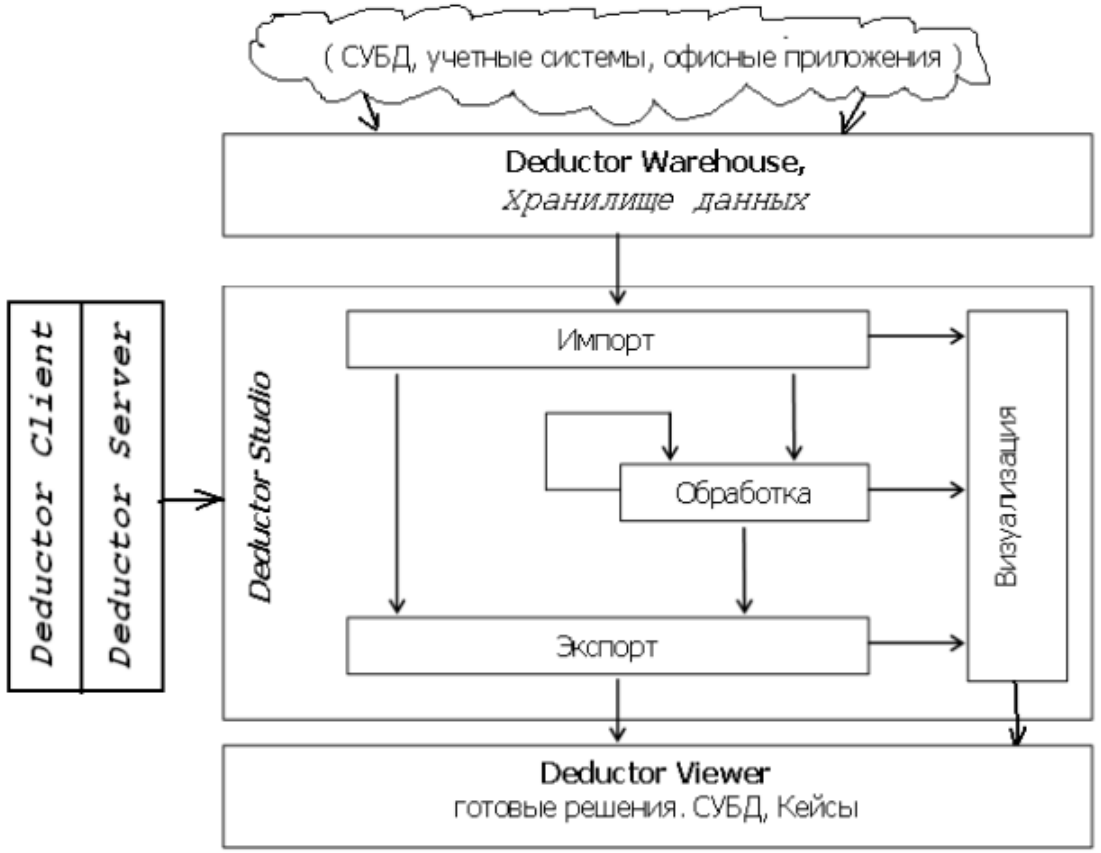
Существует три типа варианта поставки платформы Deductor:

- Enterprise;
- Professional;
- **Academic.**

Аналитическая платформа Deductor

Архитектура системы построена таким образом, что вся работа по анализу данных в Deductor Studio базируется на выполнении следующих действий:

- импорт данных;
- обработка данных;
- визуализация;
- экспорт данных.



Вопросы для проверки

1. Что такое KDD?
2. Из каких этапов состоит KDD?
3. Что такое хранилище данных?
4. Что такое Data Mining?
5. Чем отличается Machine Learning от Data Mining?
6. Каковы задачи и методы Data Mining?
7. Что такое аналитическая платформа?
8. Какие функции реализует Deductor Studio?
9. Какова архитектура АП Deductor?