ВВЕДЕНИЕ В ДИСЦИПЛИНУ "ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ"

Понимание приходит после анализа сомнений.
Г. Александров

Содержание

2

- Цель и задачи дисциплины
- Программа дисциплины
- Литература и web-ресурсы
- Обзор изучаемых тем
- Обзор курсового проекта

Цель и задачи дисциплины

3

- *Цель* – ознакомление с основами технологий хранилищ данных (Data Warehouse), оперативного анализа данных (OLAP) и интеллектуального анализа данных (Data Mining).
- Основные задачи – приобретение компетенций в следующих областях:
  - построение хранилищ данных
  - разработка OLAP-приложений
  - разработка приложений Data Mining.

Программа дисциплины

4

- Лекции (22 час.)
  - Введение
  - Предварительная обработка данных (Data Preprocessing)
  - Хранилища данных (Data Warehouse)
  - Оперативный анализ данных (OLAP)
  - Интеллектуальный анализ данных (Data Mining)
- Контрольные мероприятия
  - Экзамен (тест)
  - Курсовой проект (разработка приложения, написание отчета, защита проекта)

Литература и web-ресурсы

5

- *Барсегян А.А. и др.* Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. СПб.: БХВ-Петербург, 2007.

- *Han J., Kamber M.* Data Mining: Concepts and Techniques, 2nd ed. Morgan Kaufmann Publishers, 2006.
- Страница курса http://mzym.susu.ru/courses/olap/

"Всемирный потоп" данных

6

Единица	Значение
1 Кбайт	2^{10} байт
1 Мбайт	2^{10} Кбайт
1 Гбайт	2^{10} Мбайт
1 Тбайт	2^{10} Гбайт
1 Пбайт	2^{10} Тбайт
1 Эбайт	2^{10} Пбайт
1 Збайт	2^{10} Эбайт
1 Йбайт	2^{10} Збайт

- Бизнес: электронная коммерция, банковские транзакции, биржи и др.
- Наука: сенсорные сети, данные моделирования и др.
- Общество: новостные ленты, социальные сети, YouTube и др.

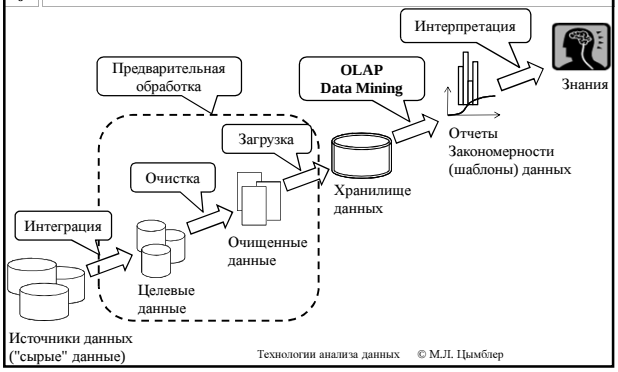
Избыток данных, недостаток знаний

7

- Проблема: как найти необходимые данные?
  - Данные разбросаны по сети
  - Данные имеют много версий с трудноуловимыми отличиями
- Проблема: как получить необходимые данные?
  - Для получения необходимых данных нужен эксперт
- Проблема: как понять найденные данные?
  - Имеющиеся данные плохо документированы
- Проблема: как использовать найденные данные?
  - Неожиданные результаты
  - Необходимость преобразования данных из одной формы в другую

Аналитическая обработка данных

8



Предварительная обработка

9

- Неполнота сырых данных
  - Примеры: "N/A" и NULL, Город=" "
  - Причины: проблемы ПО/АО/человеческий фактор
- Шумы в сырых данных
  - Примеры: Зарплата = -10000
  - Причины: ошибки человека/компьютера при сборе/передаче данных
- Несогласованность сырых данных
  - Примеры: Категория = "1, 2, 3" или "A, B, C"; Город=Челябинск и Страна=США
  - Причины: различные источники данных, плохое проектирование баз данных и/или их приложений, дублирование данных
- Качественные сырые данные ⇒ качественные знания

Предварительная обработка

10

- Очистка данных
  - Заполнить пропущенные значения, сгладить шумы в данных, определить или удалить аномалии, исправить несогласованности
- Интеграция данных
  - Интегрировать базы данных, документы различной природы
- Трансформация данных
  - Нормализовать и агрегировать
- Редукция данных
  - Получить данные меньшего объема, аналитическая обработка которых дает такие же или схожие результаты

Хранилище данных

11

- *Хранилище данных (Data Warehouse)* – предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для решения задач аналитической обработки данных.
- Основная идея – *разделение данных*, используемых для оперативной и аналитической обработки.

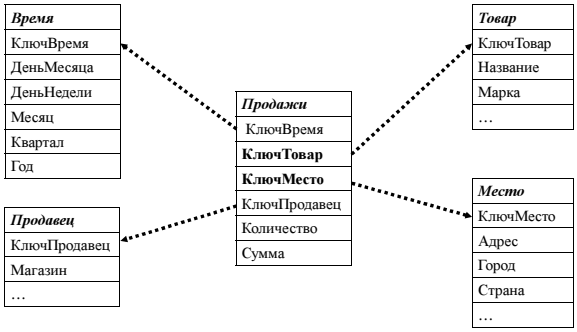
База данных *vs.* хранилище данных

12

- *База данных* – субъект OLTP *(On-line Transaction Processing, оперативная обработка транзакций)*
  - Повседневные операции: покупки, бухгалтерия, регистрация, и др.
- *Хранилище данных* – субъект *OLAP (On-line Analytical Processing, оперативный анализ данных)* и *Data Mining (интеллектуальный анализ данных)*
  - Анализ данных и принятие решений
- Отличия
  - Направленность пользователей и систем: покупатель *vs.* рынок
  - Данные: текущие, детализированные *vs.* исторические, консолидированные
  - Проектирование: ER + приложение *vs.* схема "звезда"
  - Доступ к данным: update *vs.* read-only, сложные запросы

Хранилище данных

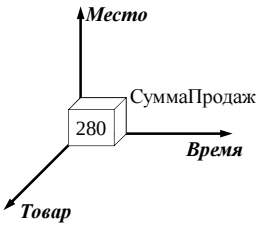
13



OLAP, многомерные данные

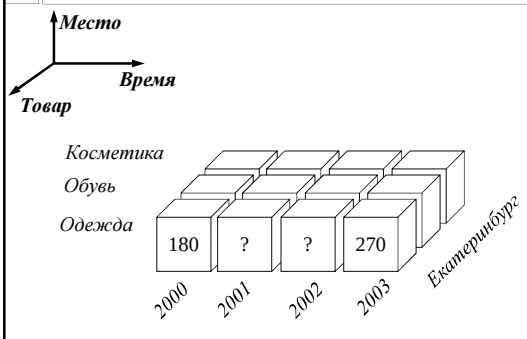
14

- *OLAP (On-line Analytical Processing)* – оперативная аналитическая обработка данных, поддерживающая многомерную модель данных.



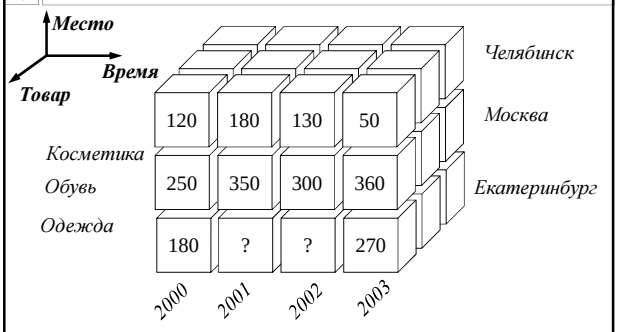
Куб данных

15



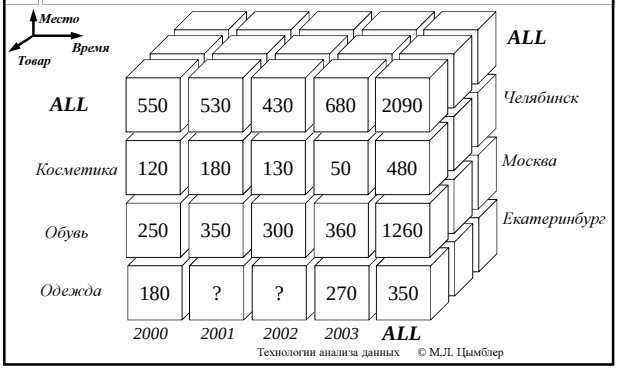
Куб данных

16



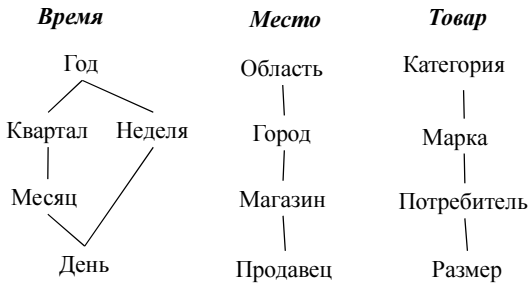
Гиперкуб (OLAP-куб)

17



Иерархия в измерениях

18



OLAP-технологии

19

- Сложные запросы выборки по большому количеству критериев и операций агрегации
  - получить список продавцов из Челябинска, продавших в 1 квартале женских пиджаков марки MEXX размера S на сумму, большую, чем средняя сумма продаж женских пиджаков в Челябинской области за этот период.
- OLAP-расширение SQL
  - select Время, Место, Товар,
    sum(СуммаПродаж) as Прибыль
    from Продажи **cube by** (Время, Место, Товар);

CUBE BY

20

Время	Филиал	Товар	Прибыль
2000	Челябинск	Одежда	100 000
2000	Челябинск	Косметика	120 000
2000	Москва	Одежда	250 000
2000	Москва	Косметика	75 000
2001	Челябинск	Одежда	230 000
2001	Челябинск	Косметика	310 000
2001	Москва	Одежда	170 000
2001	Москва	Косметика	350 000

CUBE BY

21

```
select
    Время, Место,
    Товар, sum(Прибыль)
    as Прибыль
from Продажи
cube by (Время,
        Место, Товар)
```

Время	Место	Товар	Прибыль
2000	**Челябинск**	**Одежда**	100 000
2000	Челябинск	**Косметика**	120 000
2000	**Челябинск**	**[NULL]**	**220 000**
2000	**Москва**	**Одежда**	250 000
2000	Москва	**Косметика**	75 000
2000	**Москва**	**[NULL]**	**325 000**
2000	**[NULL]**	**Одежда**	**350 000**
2000	[NULL]	**Косметика**	**195 000**
2000	**[NULL]**	**[NULL]**	**545 000**
2001	**Челябинск**	**Одежда**	230 000
2001	Челябинск	**Косметика**	310 000
2001	**Челябинск**	**[NULL]**	**540 000**
2001	**Москва**	**Одежда**	170 000
2001	Москва	**Косметика**	350 000
2001	**Москва**	**[NULL]**	**520 000**

CUBE BY

22

```
select
    Время, Место,
    Товар, sum(Прибыль)
    as Прибыль
from Продажи
cube by (Время,
         Место, Товар)
```

Время	Место	Товар	Прибыль
[NULL]	**Челябинск**	**Одежда**	330 000
[NULL]	Челябинск	**Косметика**	430 000
[NULL]	**Челябинск**	**[NULL]**	**760 000**
[NULL]	**Москва**	**Одежда**	420 000
[NULL]	Москва	**Косметика**	425 000
[NULL]	**Москва**	**[NULL]**	**845 000**
[NULL]	[NULL]	**Одежда**	**750 000**
[NULL]	[NULL]	**Косметика**	**855 000**
[NULL]	**[NULL]**	**[NULL]**	**1 605 000**

Интеллектуальный анализ данных

23

- *Интеллектуальный анализ данных (Data Mining)* – процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.
- Другие термины
  - Knowledge Discovery in Databases
  - Machine Learning
  - Business Intelligence

Сферы применения

24

- Банковская сфера, страхование
  - Оценка платежеспособности
  - Решение о выдаче кредита
- CRM (Customer Relationship Management) и маркетинг
  - Клиенты, которые с наибольшей вероятностью уйдут к конкурентам
  - Товары, наиболее часто продающихся совместно
  - Рекомендательные сервисы
- Телекоммуникации, электронные транзакции
  - Отсечение спама
  - Выявление мошенничества
  - Персонализация web-сайтов
- Медицина и фармация
  - Системы постановки диагноза
  - Анализ истории болезни, определение эффективности лечения

Основные задачи Data Mining

25

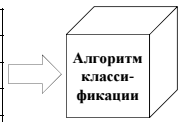
- *Классификация* – определение класса объекта по характеристикам этого объекта, когда множество классов заранее известно.
- *Прогноз* – определение значения некоторого параметра объекта (вещественного числа) по характеристикам этого объекта.
- *Кластеризация* – определение классов объектов по характеристикам этих объектов, когда множество классов заранее неизвестно.
- *Поиск ассоциативных правил* – нахождение частых зависимостей между объектами (событиями).

Классификация

26

Обучающая выборка

ФИО	**Доход**	**Возраст**	Выдать кредит
Иванов	Низкий	<30	**НЕТ**
Петров	Средний	30..40	**ДА**
Сидоров	Высокий	<30	**ДА**
Егоров	Средний	>40	**ДА**
Бендер	Низкий	30..40	**ДА**
Балаганов	Средний	<30	**НЕТ**



ЕСЛИ Доход=Высокий
ИЛИ Возраст>30
ТО ВыдатьКредит=ДА

Модельный классификатор

Классификация

27

Модельный классификатор

ЕСЛИ Доход=Высокий
ИЛИ Возраст>30
ТО ВыдатьКредит=ДА

Тестовая выборка

ФИО	**Доход**	**Возраст**	Выдать кредит
Васечкин	Низкий	<30	**НЕТ**
Петров	Средний	<30	**НЕТ**
Воробьянинов	Высокий	>40	**ДА**
Бонд	Средний	30..40	**ДА**

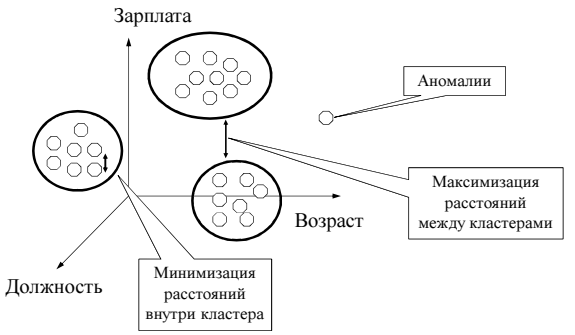
Точность модели

Выдать кредит	ОК
НЕТ	✓
ДА	×
ДА	✓
ДА	✓

75%

Кластеризация

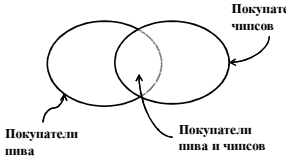
28



Поиск ассоциативных правил

29

Покупка	Товары
1	пиво, вода, чипсы
2	пиво, соль, чипсы
3	пиво, чипсы, орехи
4	вода, орехи, хлеб
5	вода, соль, чипсы, орехи, хлеб



- Найти все правила $X \rightarrow Y$, имеющие поддержку и достоверность больше, чем заданные.
  - *Поддержка* – вероятность того, что покупка содержит товары $X \cup Y$
  - *Достоверность* – условная вероятность того, что покупка, содержащая товар X, также содержит товар Y.
- Частоты
  - {пиво: 3; вода: 3; чипсы: 4; орехи: 3; пиво+чипсы: 3}
- Ассоциативные правила
  - пиво → чипсы (0,6; 1)
  - чипсы → пиво (0,6; 0,75)

Курсовой проект

30

- Разработка системы поддержки принятия решений для модельной предметной области*
  (см. http://mzym.susu.ru/courses/olap/).
- Методические указания.
- Контрольные точки.
- Подготовка и защита отчета.

* Возможны другие варианты модельной предметной области по согласованию с преподавателем.

Задание

31

- Торговая фирма
  - Продажа обуви, галантереи и одежды для детей, мужчин и женщин.
  - Три географически удаленных друг от друга магазина отдельно ведут учет своих продаж.
- Аналитик (маркетолог)
  - определение наиболее продаваемых товаров по различным параметрам
  - определение наборов товаров, часто продаваемых совместно
  - …

Задание: Магазин #1

32

- Продавцы
  (ТабНомер, ФИО, Пол)
- Товары
  (Артикул, Название, Цена)
  - Артикул=*aa-bb-ccccc-ddddd*
    - *aa*
      - OD=Одежда
      - GA=Галантерея
      - OB=Обувь
    - *bb*
      - 01=Дети
      - 02=Мужчины
      - 03=Женщины
    - *ccc*
      - WHITE
      - RED##
      - BLACK
      - GREEN
      - BLUE#
- Продажи (Чек, Артикул^, Продавец^, Количество, Дата)

Задание: Магазин #2

33

- Продавцы
  (КодПродавца, ФИО, Пол)
- Продажи
  (КодЧека, НазваниеТовара, ВидТовара, Потребитель, ЦветТовара, ЦенаТовара, КодПродавца^, Количество, Дата)
  - ВидТовара: 'одежда', 'галантерея' и 'обувь'.
  - Потребитель: 'Д', 'М' и 'Ж'.
  - ЦветТовара: 'белый', 'красный', 'черный', 'зеленый' и 'синий'.

Задание: Магазин #3

34

- Продавцы
  (ИдПродавца, ФИО, Пол)
- Обувь
  (ИдОбувь, Название, Цвет, Потребитель, Цена)
  - Цвет: 'white', 'red', 'black', 'green' и 'blue'.
  - Потребитель: 'K', 'M' и 'F'.
- ПродажиОбувь
  (ИдЧека^, ИдОбувь^, ИдПродавца^, Количество)
- Галантерея
  (ИдГалантерея, Название, Цвет, Потребитель, Цена)
- ПродажиГалантерея
  (ИдЧека^, ИдГалантерея^, ИдПродавца^, Количество)
- Одежда
  (ИдОдежда, Название, Цвет, Потребитель, Цена)
- ПродажиОдежда
  (ИдЧека^, ИдОдежда^, ИдПродавца^, Количество)
- Покупки
  (ИдЧека, Дата)

Задание: работы

35

- Создание тестовых баз данных
- Разработка хранилища данных
- Разработка прототипа системы поддержки принятия решений
- Разработка OLAP-функциональности
- Разработка Data Mining-функциональности