大数据导论
Introduction to Big Data

第4.1讲: 数据的结构化表示

叶允明

计算机科学与技术学院

哈尔滨工业大学（深圳）

目录

结构化数据

- 数据是数据对象的集合
- 数据对象用一组刻画对象基本特征的属性描述
  - 对象也叫做记录、数据点、案例、样本、观测或实体等
- **属性（attribute）**是客观事物的性质或特性的计算机表示，而数据对象是由属性集合构成的
  - 例如：眼球颜色因人而异，物体的温度随时间而变。
  - 属性也叫做变量、字段、特征或维。

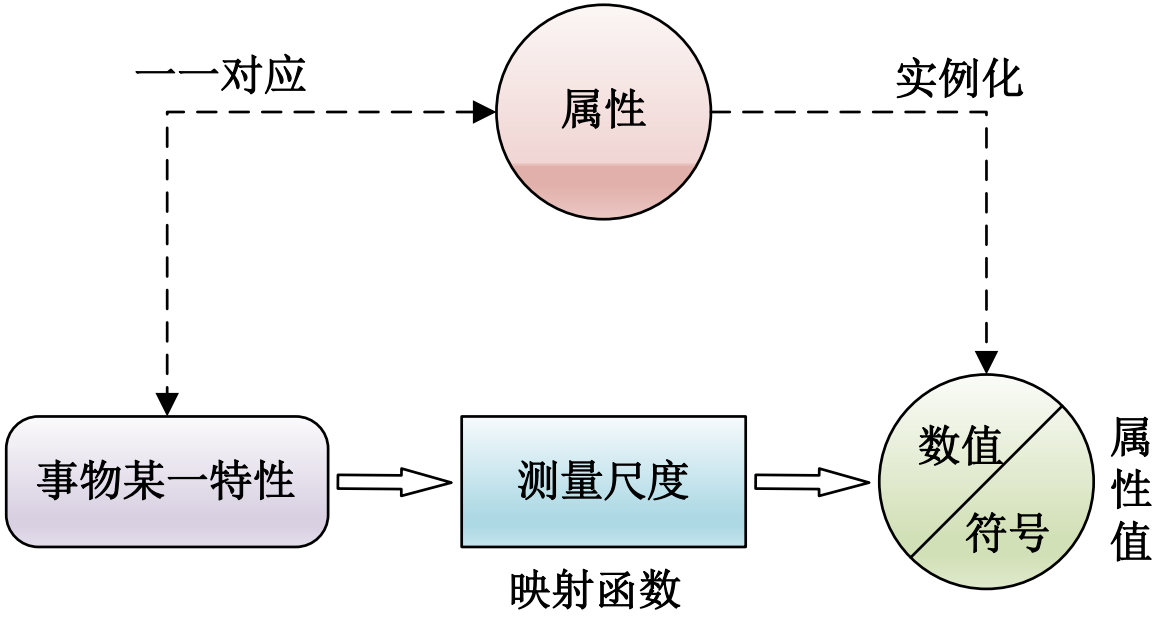
属性

对象

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Ref. this table is from Han's book slides

属性值



属性类型

- 类别型属性（categorical attribute）

  属性值定义域是一个固定、有限的符号或数字集合。

  Domain(“性别”)={M，F}

  Domain(“职业”)={教师，工程师，医生}

  属性值间不可以做算术运算：定性属性

属性类型

- 类别型属性（categorical attribute）的分类
  - 标称型属性（nominal attribute）：无序

身份证号、性别、职业、颜色

  - 有序型类别属性（ordinal attribute）：有大小、好坏等先后顺序的区别

学位：{“学士”、“硕士”、“博士”}

属性类型

- 独热编码（one-hot encoding）

职业-售货员	职业-教师	职业-白领
1	0	0
0	1	0
0	0	1

属性类型

- 数值型属性（numeric attribute）

  属性值定义在实数集或整数集上 。

  年龄、每月工资、债务收入比

  属性值间可以做算术运算：定量属性

属性类型

- 数值型属性（numeric attribute）
  - 区间标度型属性（nominal attribute）：有序

“摄氏温度”　“30℃比10℃高20℃” ✓　属性值“0℃”并不代表无温度 ✓

“30℃的温度是10℃的三倍” ✗

  - 比率标度型属性（nominal attribute）：有序

“每月工资”　月薪1万元比月薪5000元多5000 ✓　月薪为0是有实际意义的 ✓

月薪1万是月薪5000的两倍 ✓

离散属性和连续属性

- 离散属性
  - 具有有限个值或无限可数个值
  - 例：邮政编码，计数或文档集合中的单词集
  - 通常表示为整数变量
  - 注：二元属性是离散属性的一种特殊情况
- 连续属性
  - 将实数作为属性值
  - 例：温度，高度，或重量
  - 实践中，实数值只能用有限的精度测量和表示
  - 连续属性通常用浮点变量表示

数据的结构化

- “**数据的结构化**”就是指将原始数据按照固定的“属性-值”序列逐行排列各个数据对象，其结果就形成了“**结构化数据**”。

- 标准结构化数据的形式化定义：

$$\boldsymbol{D} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_i}, \ldots, \boldsymbol{x_m}\},\ (i = 1 \ldots m),$$

$$\text{其中}\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in}\}\,(j = 1 \ldots n),\ x_{ij} \in \boldsymbol{R}$$

单一类型数据的结构化表示

关系表数据的结构化表示

- 单个关系表数据的结构化
  - 关系表数据是指以二维表形式存储的结构化数据

用户 ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	白领

关系表数据的结构化表示

- 单个关系表数据的结构化

用户 ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	白领

用户 ID	性别	年龄	月薪	职业-售货员	职业-教师	职业-白领
1	1	45	5000	1	0	0
2	0	35	10000	0	1	0
3	1	28	9000	0	0	1

关系表数据的结构化表示

- 多个关系表数据的结构化

客户基本信息表

用户ID	性别	年龄	月薪	职业
1	男	45	5000	售货员
2	女	35	10000	教师
3	男	28	9000	白领

客户借贷记录信息表

用户ID	累计借款金额	累计借款天数	是否存在逾期未还
1	20000	365	1
2	8000	30	0
3	15000	90	0

用户ID	性别	年龄	月薪	职业-售货员	职业-教师	职业-白领	累计借款金额	累计借款天数	是否存在逾期未还
1	1	45	5000	1	0	0	10000	365	1
2	0	35	10000	0	1	0	8000	30	0
3	1	28	9000	0	0	1	15000	90	0

文本数据的结构化表示

- 标准结构化数据——数值矩阵
  - 文本数据

    高维、稀疏

  - 结构化表示：词频法

Text_1:“背包设计的很好看，质量很好，价格便宜”

Text_2: “这是我买过性价比最高的背包，很好看”

Text_1: “背包\设计\的\很\好看\，\质量\很\好\，\价格\便宜”

Text_2: “这\是\我\买\过\性价比\最高\的\背包\，\很\好看\”

文本数据的结构化表示

- 词频法

Text_1: “背包\设计\的\很\好看\，\质量\很\好\，\价格\便宜”

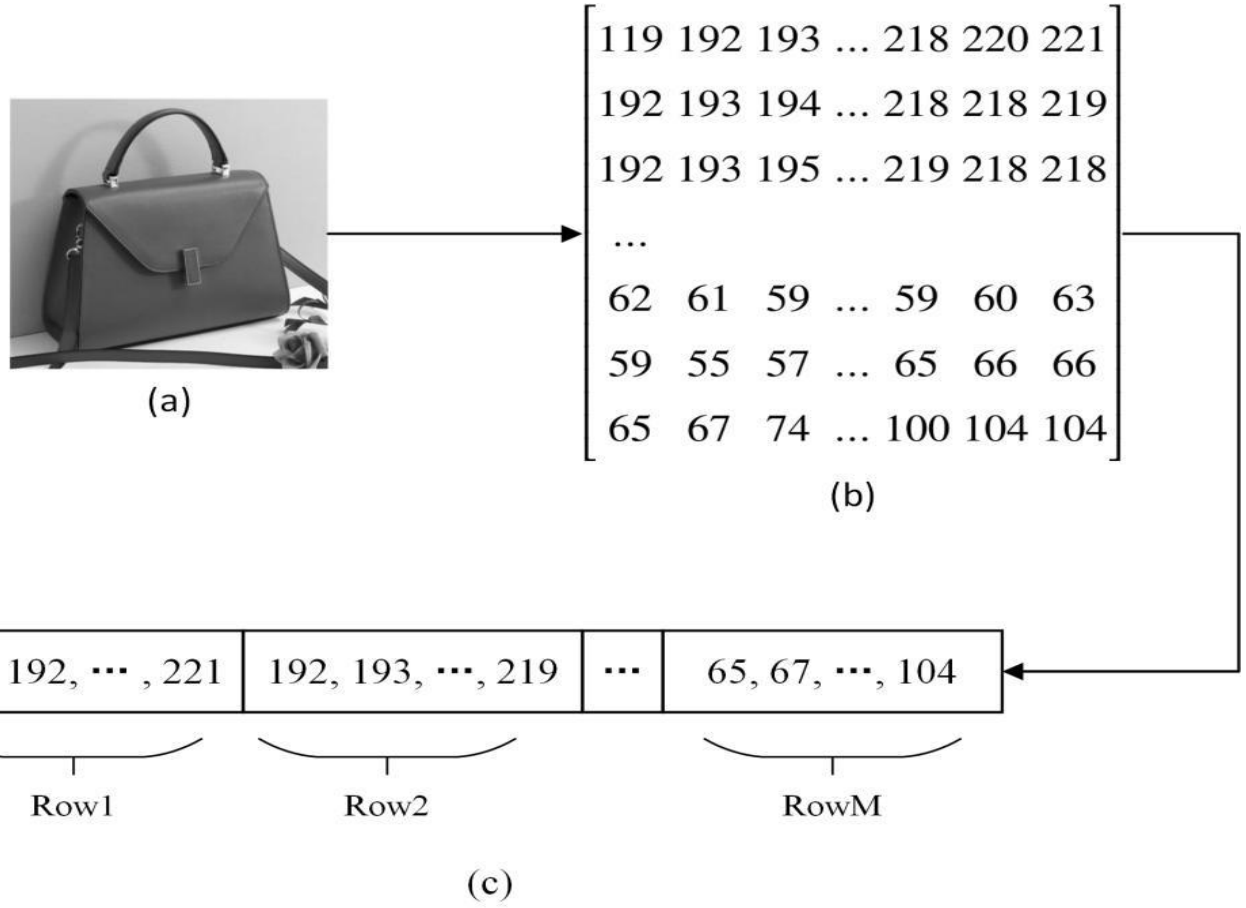
Text_2: “这\是\我\买\过\性价比\最高\的\背包\，\很\好看\”

统计词频

词语 文本	很	的	好	这	是	我	买	过	背包	设计	好看	质量	价格	便宜	最高	性价比
Text_1	2	1	1	0	0	0	0	0	1	1	1	1	1	1	0	0
Text_2	1	1	0	1	1	1	1	1	1	0	1	0	0	0	1	1

图像数据的结构化表示

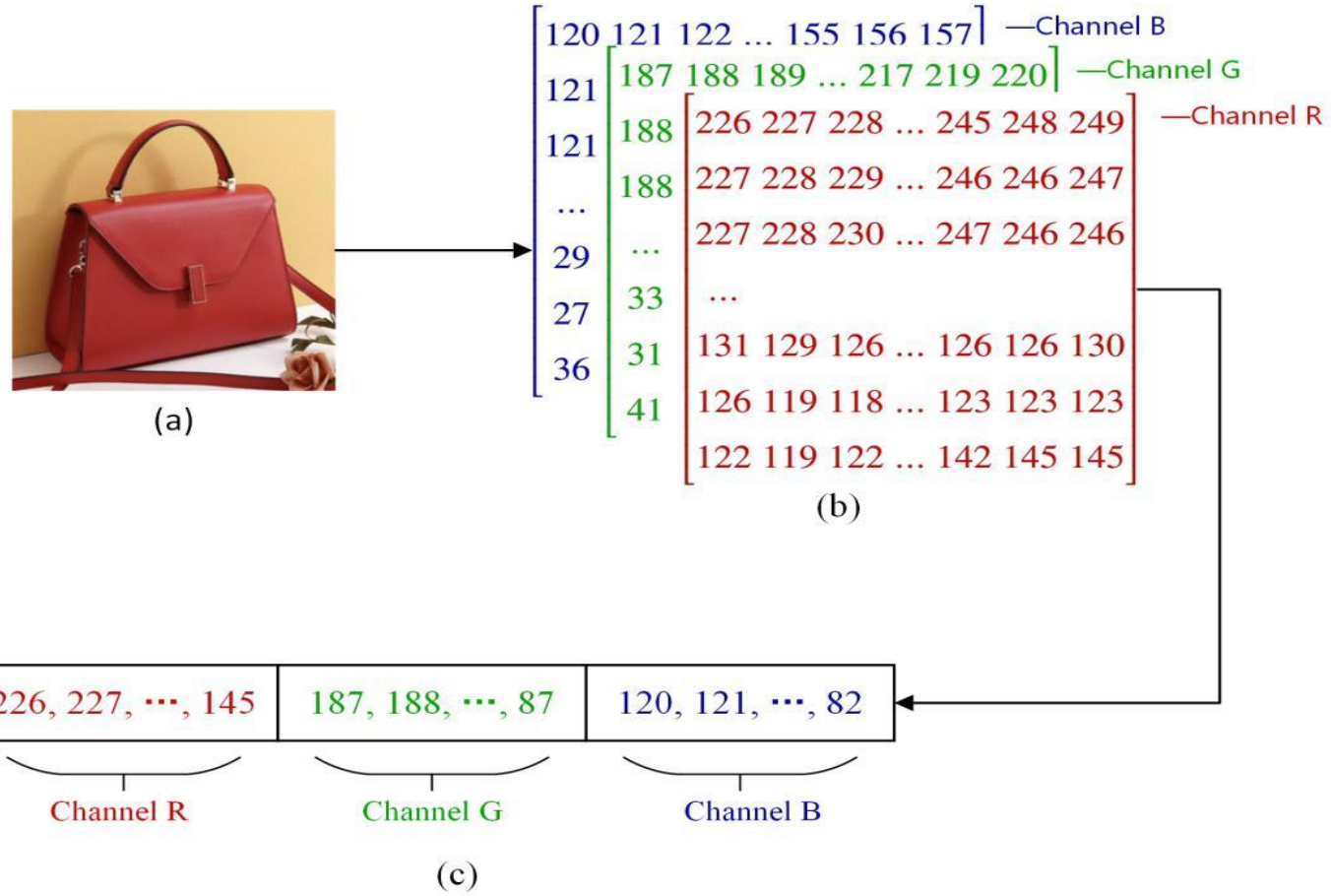
- 标准结构化数据——数值矩阵
  - 数字图像与视频数据
    - ✓ 结构化表示：扁平化



单通道灰度图像的结构化表示

图像数据的结构化表示

- 标准结构化数据——数值矩阵
  - 数字图像与视频数据
    - ✓ 结构化表示：扁平化



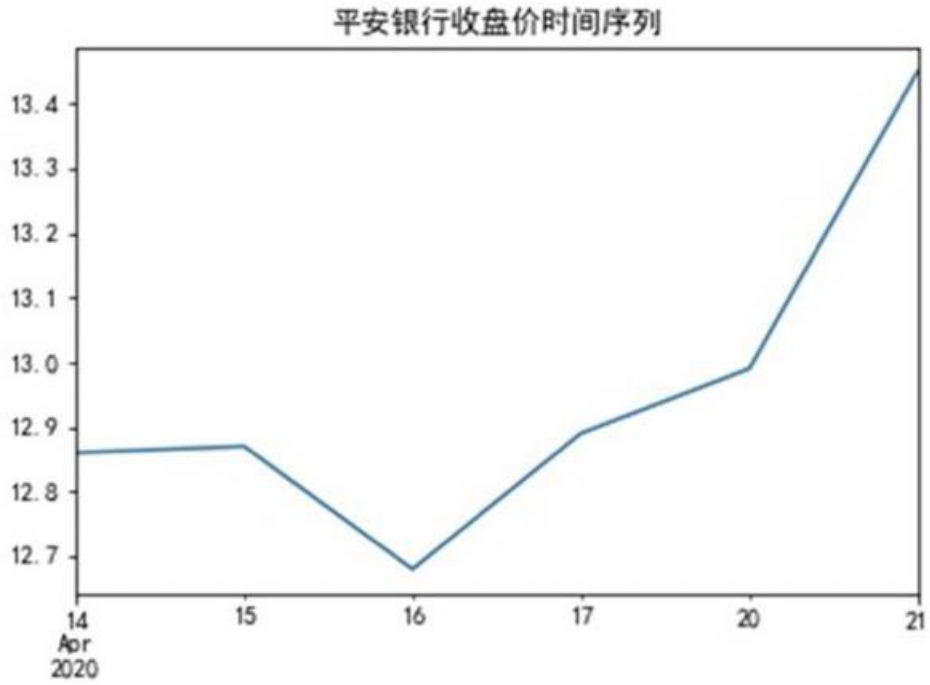
多通道彩色图像的结构化表示

时序数据的结构化表示

- 时间序列数据（time series data），是在不同的时间间隔观测同一数据对象，将其属性值相继排列所形成的序列
- 单通道时序数据

trade_date	close
2020/4/14	12.86
2020/4/15	12.87
2020/4/16	12.68
2020/4/17	12.89
2020/4/20	12.99
2020/4/21	13.45

(a)



(b)

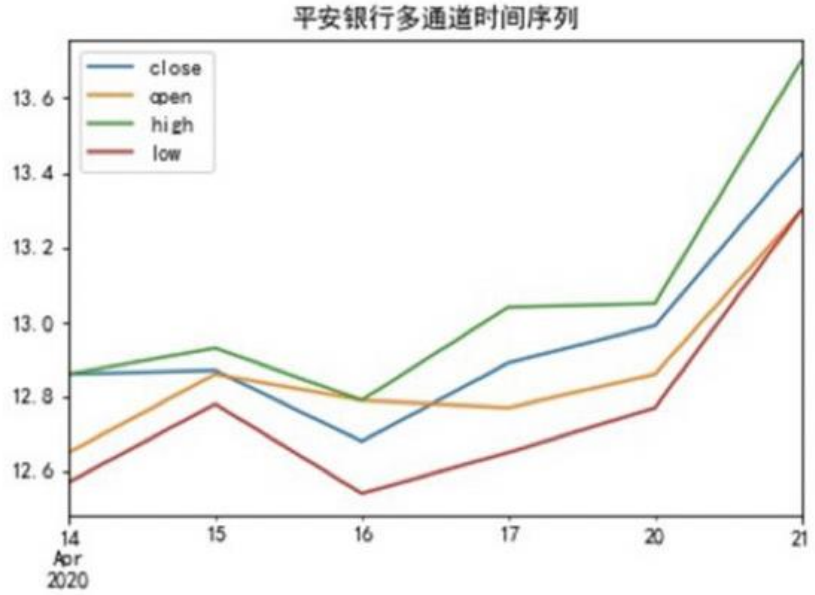
单通道：股票的"每日收盘价"的涨跌情况

时序数据的结构化表示

- 多通道时序数据

trade_date	close	open	high	low
2020/4/14	12.86	12.65	12.86	12.57
2020/4/15	12.87	12.86	12.93	12.78
2020/4/16	12.68	12.79	12.79	12.54
2020/4/17	12.89	12.77	13.04	12.65
2020/4/20	12.99	12.86	13.05	12.77
2020/4/21	13.45	13.3	13.7	13.3

(a)



(b)

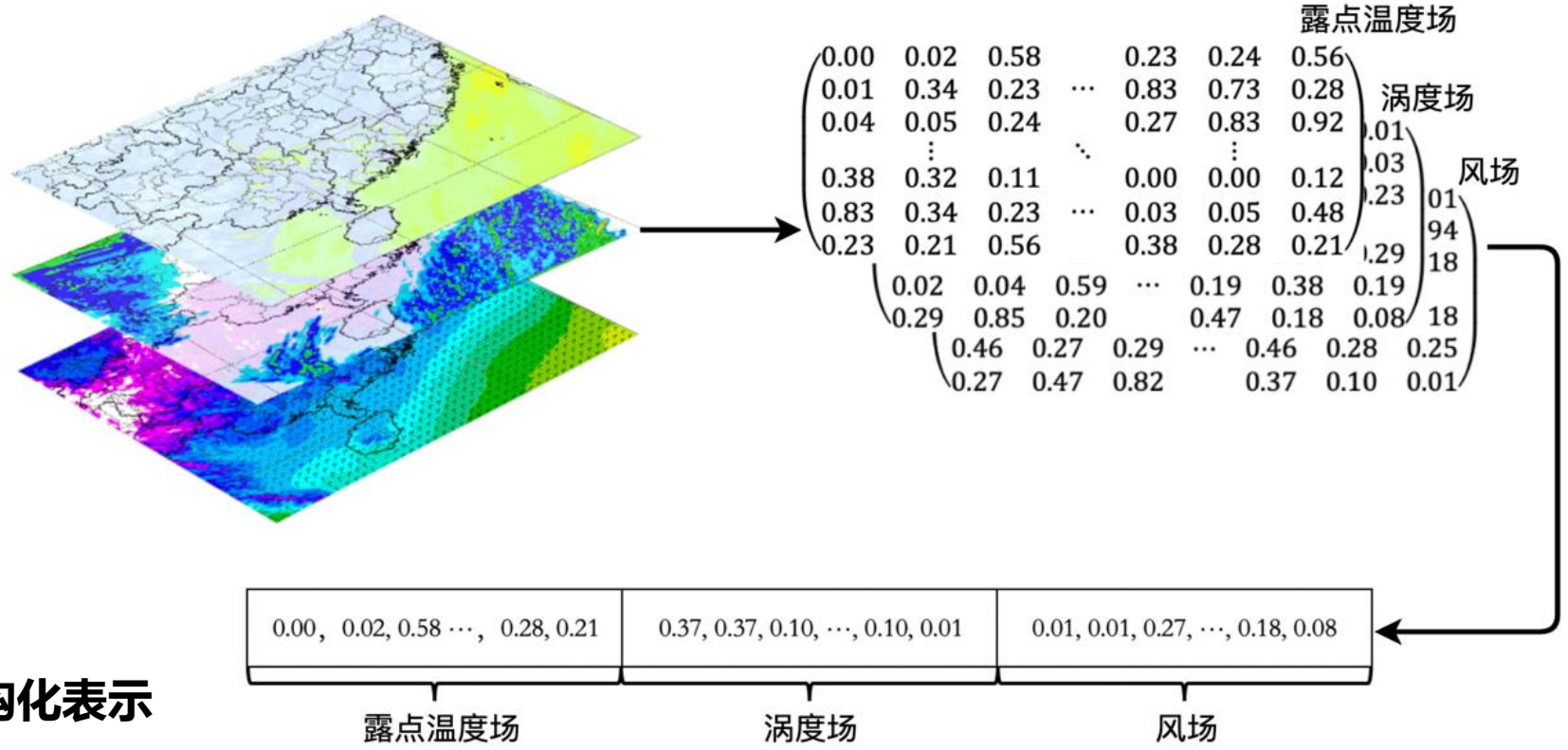
多通道：股票的“每日开盘价”、“每日收盘价”、“每日最高价”、“每日最低价”涨跌情况

空间数据的结构化表示

- 空间数据：用于描述现实世界中空间物体的位置、形态、大小、分布等特征信息的数据。
  - ✓ 非对象模型 / 矢量模型 （点、线、面 ）
  - ✓ 场模型 / 栅格模型 （二维矩阵或张量（高维矩阵） ）

空间数据的结构化表示

- 空间数据的结构化表示：先将空间场数据网格化为数值矩阵数据，再将矩阵数据逐行拼接得到向量。



空间场数据的结构化表示

多源异构数据的统一结构化表示

多源异构数据的结构化表示

- 输入数据来自多个来源、且包含多种类型——多源异构
  - 电商平台商品销售额预测
    - ✓ 商品的基本属性数据：商品ID、商品的大类、商品单价、颜色、尺寸等，属于关系表类型的数据
    - ✓ 商品描述数据：对每个商品（对应于每个商品ID）的文字描述，属于文本数据；
    - ✓ 商品的展示图片：每个商品（对应于每个商品ID）的外观图，属于图像数据。

多源异构数据的结构化表示

商品的展示图片

商品描述数据

"2020新款红色时尚牛皮包，洋气百搭"

商品的基本属性数据

商品ID	商品价格	商品类别
312	99.9	服饰

广告ID	商品ID	开始时间	结束时间
12	312	2020/5/31 17:00	2020/5/31 21:00
25	312	2020/6/1 17:00	2020/6/1 21:00

扁平化图像矩阵 → [226, 227, 228, ···, 76, 80]
向量维度:360*360*3

词频法 → [0, 2, 2, 0, 0, 0, 1, 0, ···, 1, 0]
向量维度:56857(总词数)

多元关系表拼接 → [99.9, 0, 1, 0, 0, 0, 8]
向量维度:7

拼接 → [226, 227, ···, 0, 2, 2, ···, 99.9, ···, 8]
向量维度:360*360*3 + 56857 + 7

原始数据

各自应用结构化表示方法
生成子向量

子向量拼接
成异构数据统一表示向量

致谢

- 一小部分图表、文字参考教材、互联网资料，仅供公益性的学习参考，在此表示感谢！如有版权要求请联系：yym@hit.edu.cn，谢谢！