

第2章 数据采集与治理

授课老师： 刘耿耿

联系方式： 13950363682

Email: 329717501@qq.com
liugenggeng@fzu.edu.cn

1

概述

2

大数据的来源与多源数据采集方式

3

数据集成和跨界应用的数据集成方法

4

数据的预处理

5

习题

概述

- 关于数据进入数据库之前的故事...
- 三个内容
 - ✓ 数据获取
 - ✓ 数据集成
 - ✓ 数据预处理
- 知识点
 - ✓ 大数据的来源
 - ✓ 大数据的获取手段
 - ✓ 数据离散化
 - ✓ 数据集成相关理论与方法
 - ✓ 数据变换
 - ✓ 数据质量

●课程重点

- ✓ 重点1 大数据的不同来源
- ✓ 重点2 不同种类大数据的采集方法以及离散化的动机
- ✓ 重点3 数据集成的概念
- ✓ 重点4 数据预处理的必要性和基本技术
- ✓ 重点5 数据质量的相关概念

●课程难点：

- ✓ 难点1 不同大数据采集方法的对象和考虑因素
- ✓ 难点2 传统数据集成和跨界数据集成的区别
- ✓ 难点3 不同数据清洗方法针对的错误类型

1

概述

2

大数据的来源与多源数据采集方式

3

数据集成和跨界应用的数据集成方法

4

数据的预处理

5

习题

世上本没有数据，一切数据都是人为的

对现实世界的测量

- 通过感知设备获得数据

人类的记录

- 由人录入计算机形成数据

计算机生成的数据

- 计算机通过现实世界模拟等程序生成数据

大数据的来源

现实世界

规模极大

更新极快

质量参差不齐

语义较为明确

价值密度较低

人类记录

规模较大

更新较快

质量很低

语义不明确

价值密度很低

计算机生成

规模可控

速度可控

质量很高

语义明确

价值密度不定

多源数据的采集

- **数据采集**是指从真实世界对象中获得原始数据的过程。
- 数据采集的过程要充分考虑其产生主体的物理性质，同时要兼顾数据应用的特点。

限制因素

- 资源有限

目标

- 有价值数据最大化
- 无价值数据最小化
- 和现实对象的偏差最小化

特殊要求

- 可靠性
- 时效性

常用的数据采集方法

01

用于采集物理世界信息的传感器

02

用于采集数字设备运行状态的日志文件

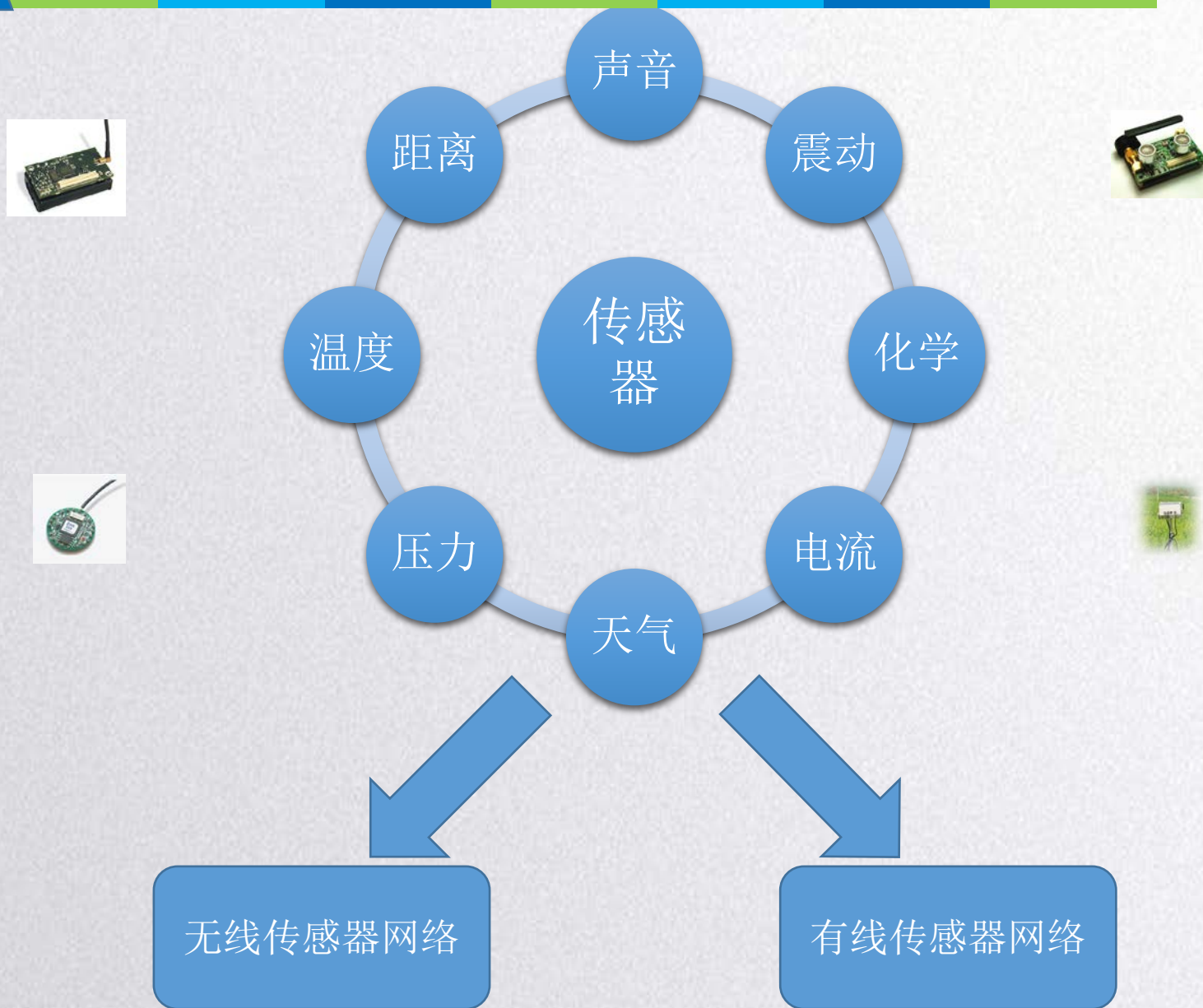
03

用于采集互联网信息的网络爬虫

04

用于采集人所了解信息的众包和群智感知技术

传感器



常用的数据采集方法

01

用于采集物理世界信息的传感器

02

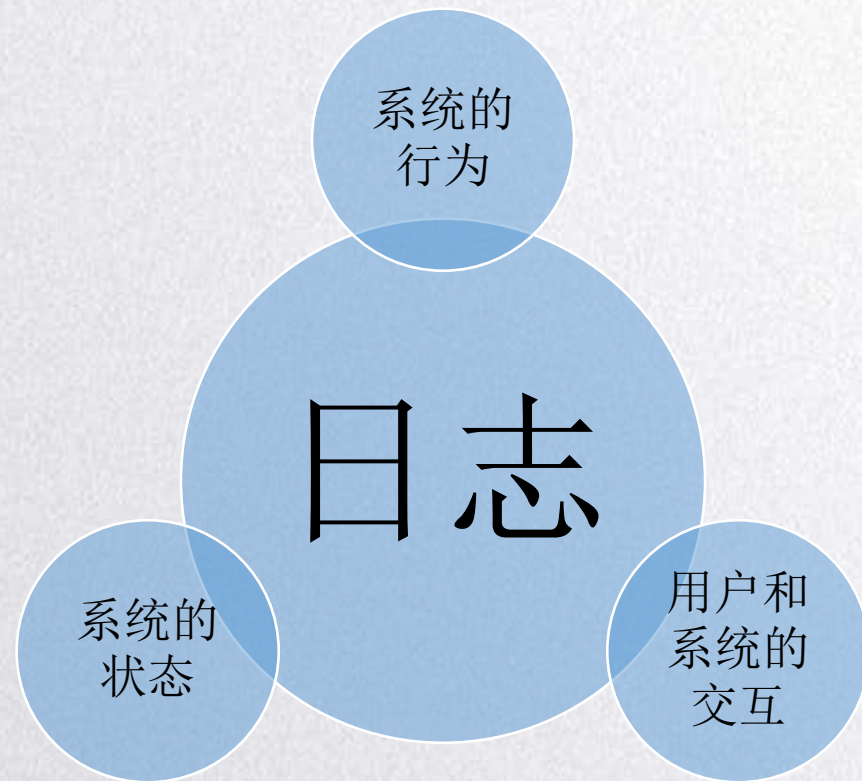
用于采集数字设备运行状态的日志文件

03

用于采集互联网信息的网络爬虫

04

用于采集人所了解信息的众包和群智感知技术



诊断系
统错误

优化运
行效率

发现用
户偏好

常用的数据采集方法

01

用于采集物理世界信息的传感器

02

用于采集数字设备运行状态的日志文件

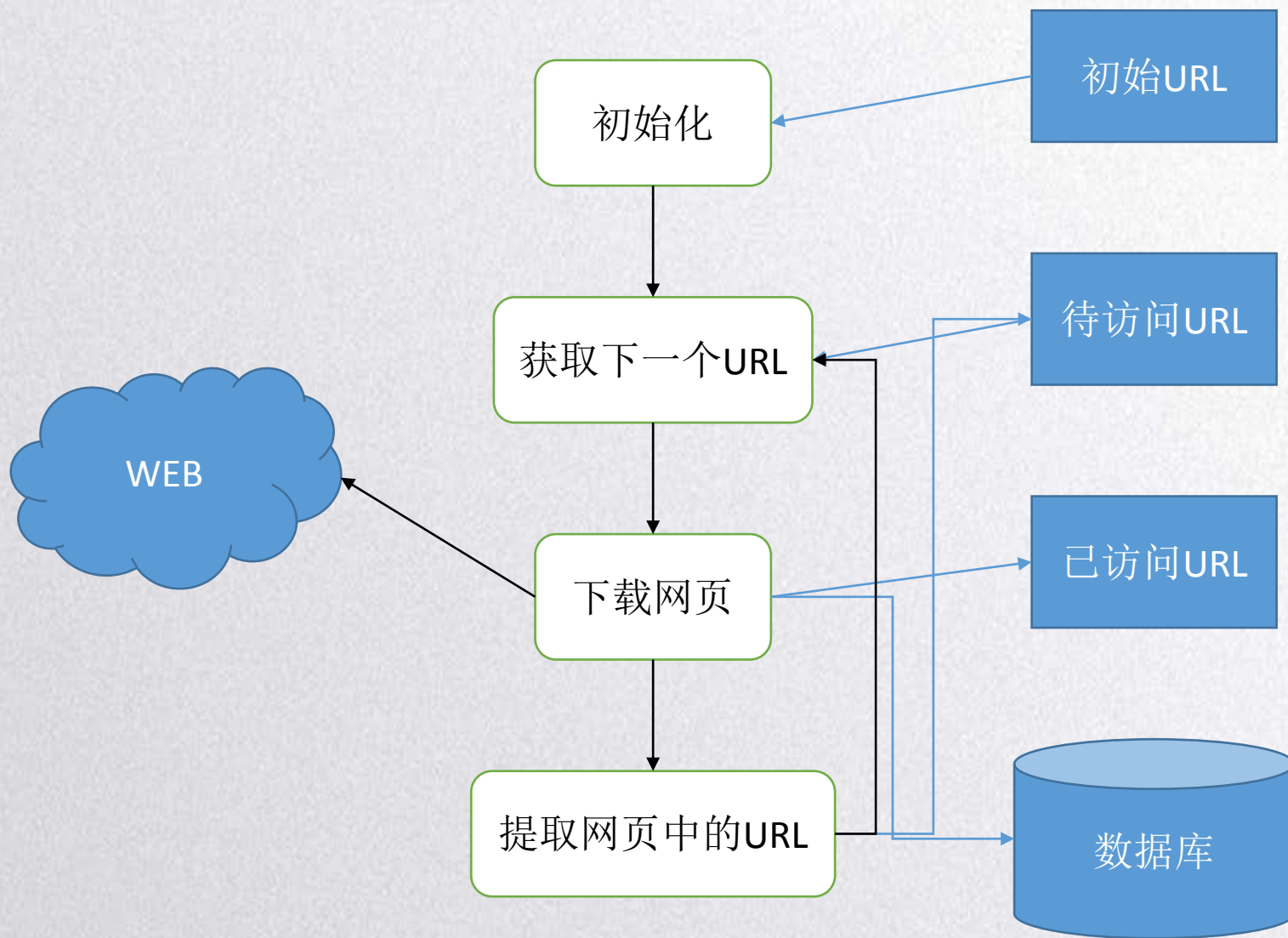
03

用于采集互联网信息的网络爬虫

04

用于采集人所了解信息的众包和群智感知技术

网络爬虫



常用的数据采集方法

01

用于采集物理世界信息的传感器

02

用于采集数字设备运行状态的日志文件

03

用于采集互联网信息的网络爬虫

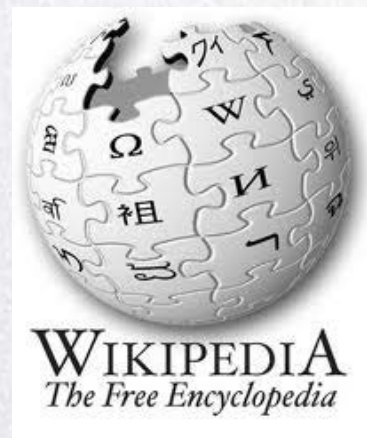
04

用于采集人所了解信息的众包和群智感知技术

众包是什么

- Outsourcing – 外包
 - 已知的雇员
- Crowdsourcing – 众包
 - 一群不固定，通常数量很大的参与者
 - 将“开源”的思想应用于软件之外

最成功的应用：Wikipedia



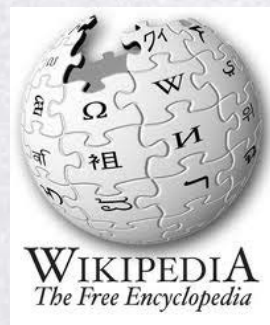
众包与群智感知

- 协调一个群体（互联网上的一大群人）做“微工作”（每人做一点贡献）来解决软件或者单个人难以解决的问题
- 通过一系列的机制和方法来指导和协调群体的行为，从而达到目的
- 群智感知：普通用户的移动设备作为基本感知单元，通过网络通信形成感知网络，从而实现感知任务分发与感知数据收集，完成大规模、复杂的社会感知任务

volunteer

fun

social



宽泛的定义

交通拥堵情况感知：一位艺术家将一百台左右智能手机都安装并开启谷歌地图

数据离散化

现实世界是连续的，因而很多传感设备采集到的都是连续的数据，而计算机只能处理以0-1形式存在的离散数据，将连续数据变成计算机可以处理的离散数据需要**数据离散化**技术。

等距

- 将连续型变量的取值范围均匀划成n等份，每份的间距相等。

等频

- 把观察点均匀分为n等份，每份内包含的观察点数相同。

优化离散

- 把自变量和目标变量联系起来考察。切分点是导致目标变量出现明显变化的折点。

当营销的重点是19-24岁的大学生群体时，可以通过优化离散技术将这部分单独划出来

客户收入属性income排序后的值(人民币元):

800 1000 1200 1500 1500 1800 2000 2300 2500 2800 3000 3500 4000 4500 4800 5000

排序后（课堂题目）

800 1000 1200 1500 1500 1800 2000 2300 2500 2800 3000 3500 4000 4500 4800 5000

01 概述

02 大数据的来源与多源数据采集方式

03 数据集成和跨界应用的数据集成方法

04 数据的预处理

05 习题

数据集成的定义与形式

数据集成是把不同来源、格式、性质的数据在逻辑上或物理上有机地集中，通过一种一致的、精确的、可用的表示法，对同一种现实世界中的实体对象的不同数据做整合的过程，从而提供全面的数据共享，经过数据分析挖掘产生有价值的信息。
可以分为**传统数据集成**和**跨界数据集成**。

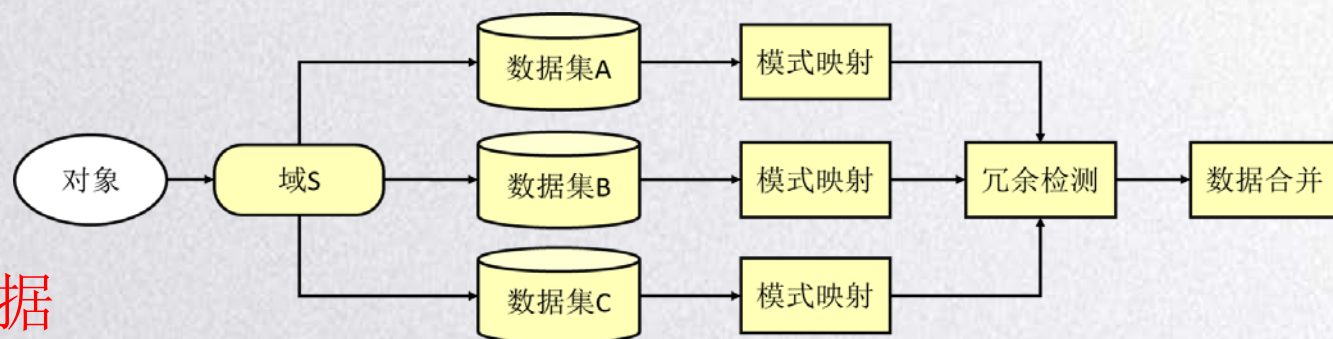


图1(a) 传统数据集成

气象数据
交通数据

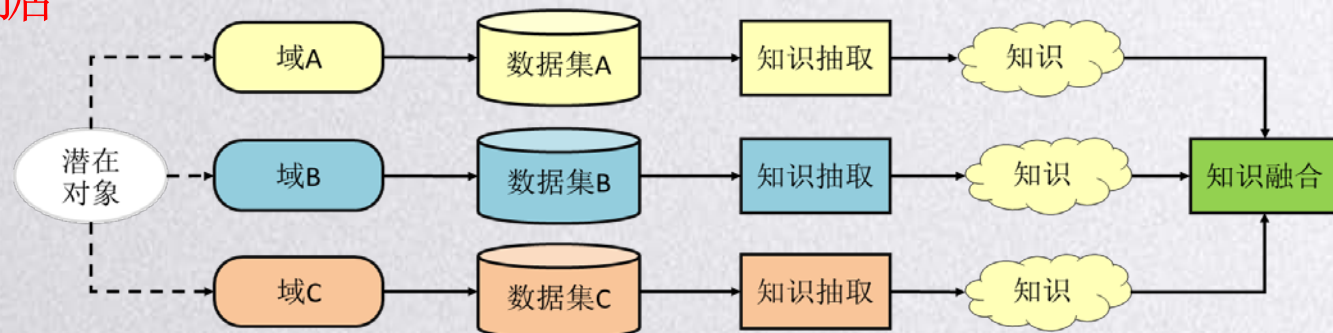
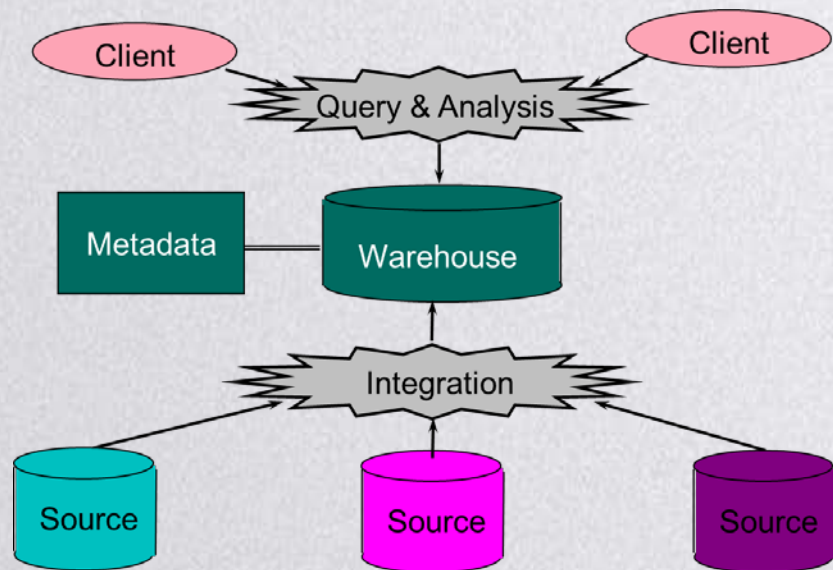


图1(b) 跨界数据集成

传统数据集成

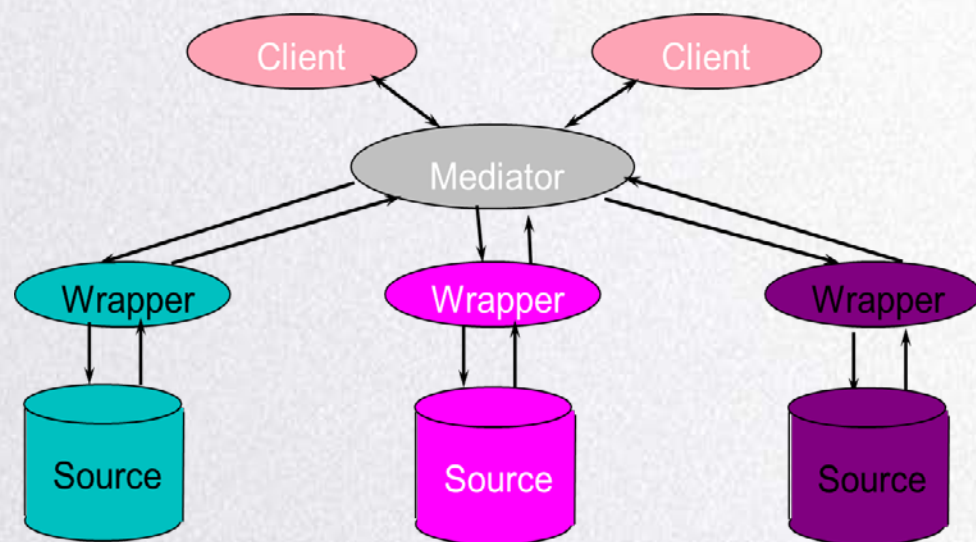
- 主要目的是数据的共享
- 定义为一个三元组 $\langle G, S, M \rangle$ ，其中：G是全局模式，S是数据源模式，M为全局模式和数据源模式之间的映射

不同数据源汇总存在单一数据库



数据仓库

Mediator不存储任何自己的数据



Mediator

模式匹配

模式匹配是标识两个数据对象是语义相关的过程

语法异构：元素语法的差异（用C语言、Java语言之间写程序的语法差异、中英文语法差异）

结构异构：元素类型、结构的差异（出生年月，出生年月日，年龄）

模型/表示异构：数据模型或其表示方法的差异（关系型数据库中一行来表示一个人的各种属性，也可以用图来表示，也可以用文档XML来表示）

各自对应什么数据结构

语义异构：同一个真实世界实体使用不同的术语描述（苹果，又称为平安果、智慧果）

数据映射

数据映射是数据在两个不同的数据模型之间进行转换的过程



数据源和目标之间的数据转换或数据中介

确定数据关系作为数据世系分析的一部分

发现隐藏的敏感数据

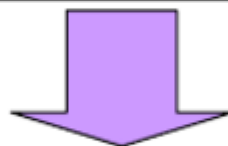
将多个数据库合并成一个数据库，并确定冗余的数据列以便合并或消除

数据映射

数据映射是

的过程

LAST_NAME	SSN	SALARY
AGUILAR	203-33-3234	40,000
BENSON	323-22-2943	60,000
D' SOUZA	989-22-2403	80,000
FIORANO	093-44-3823	45,000



LAST_NAME	SSN	SALARY
ANSKEKSL	111-23-1111	40,000
BKJHHEIEDK	111-34-1345	60,000
KDDEHLHESA	111-97-2749	80,000
FPENZXIEK	111-49-3849	45,000

列以便合并或

社会保障号404-30-5698替换为###-##-5698

数据映射

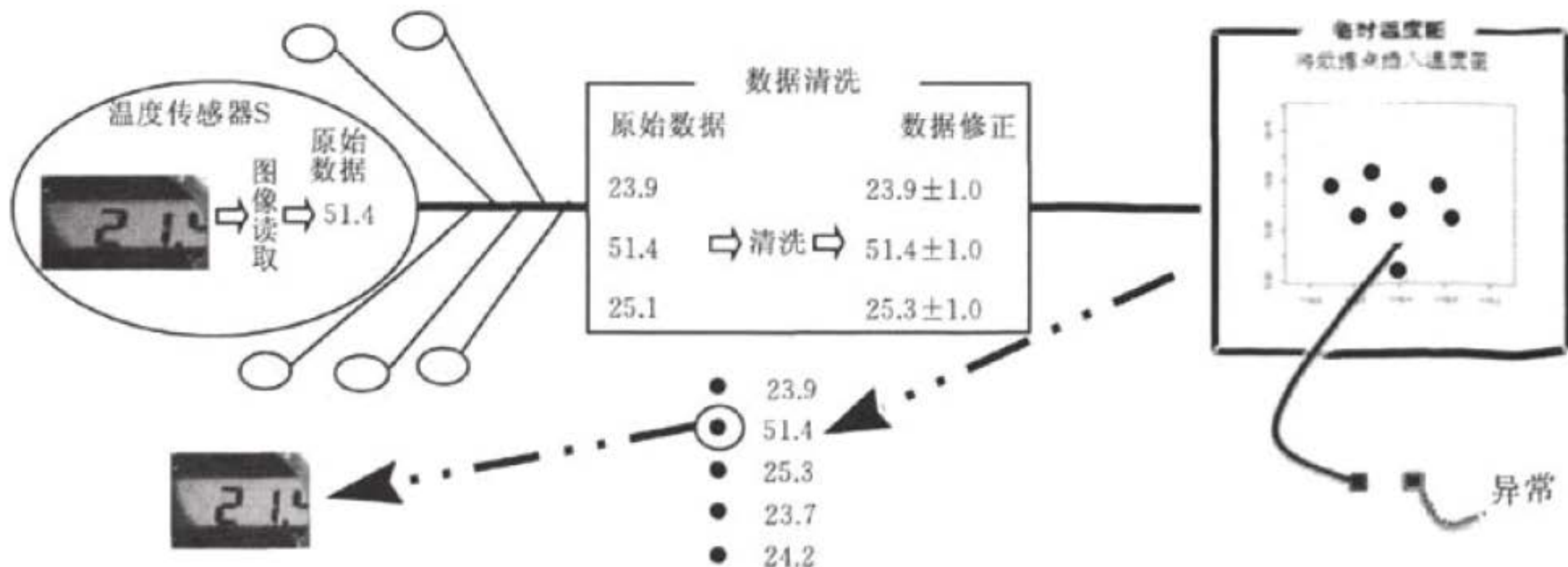



图 2 数据流世系的例子

看图说话，讲什么故事？

语义翻译

语义翻译是使用语义信息来帮助将一个数据模型中的数据转换为另一个表示或数据模型的过程

语义翻译要求源系统和目标系统中的数据元素具有到中央注册表或数据元素注册表的“语义映射”

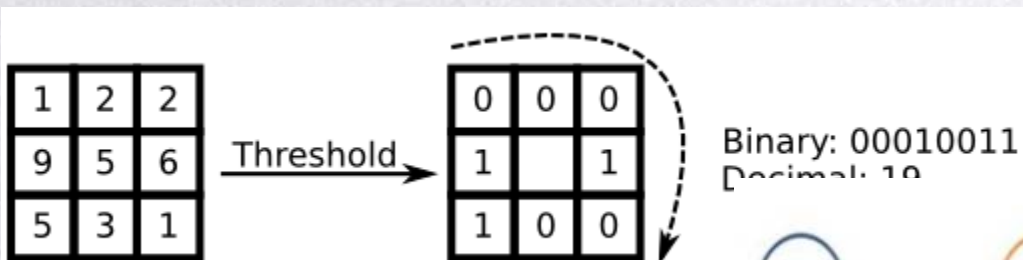


类别等价 - 表明类别或“概念”是相同的
(“人”和“个人”相同)

属性等价 - 表明两个属性是相同的
(“家庭地址”和“家庭住址”相同)

实例等价 - 表示对象的两个单独实例是等价的
(“汤姆”和“Tom”是同一个人)

跨界数据集成



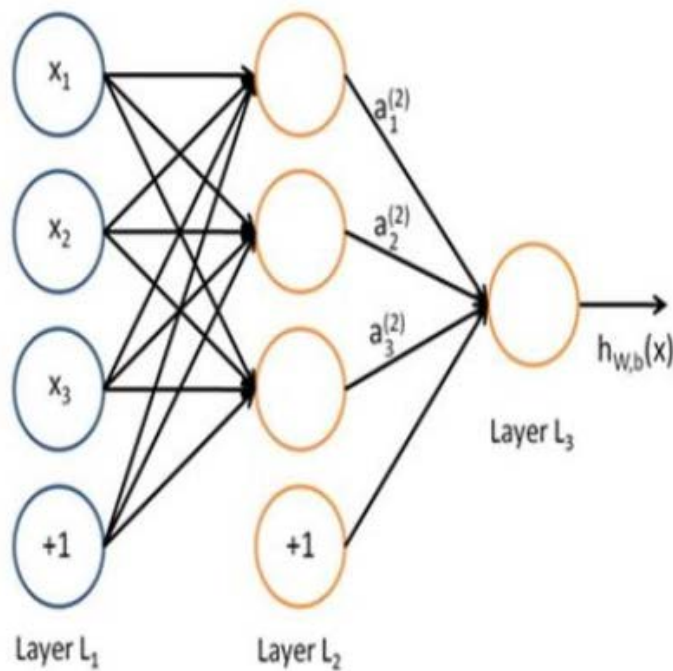
跨

基于阶段

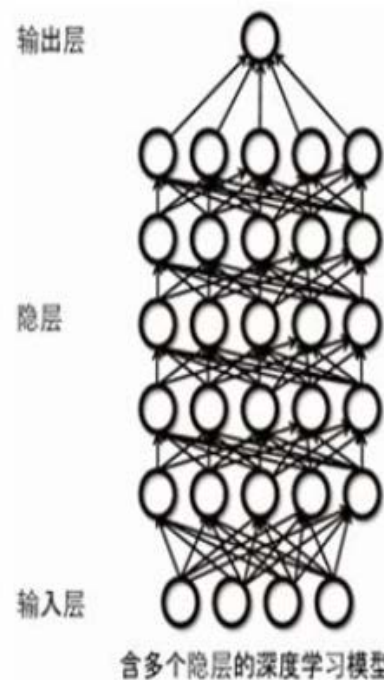
基于特征

直接关联

基于深度神经网络的方法



传统神经网络



深度神经网络

01 概述

02 大数据的来源与多源数据采集方式

03 数据集成和跨界应用的数据集成方法

04 数据的预处理

05 习题

什么是数据质量

- 如果数据适用于在操作、决策制定和计划中的角色，则其看做是高质量的
- 如果数据正确描述其指示现实世界中的对象，则称其为高质量的
- 数据质量可以有多个角度的描述方法。

精确性	一致性	完整性	时效性	实体同一性	可访问性
重复性	数据规范	表述能力	一致性表示	声誉	无害性
适量数据	安全性	可信性	易懂性	客观性	关联性
有效性	易解释性	易操作性	无误性	易用性和可维护性	使用性
可靠性	数据量	新鲜度	附加价值	易学习性	数据衰败
简洁度	一致性和	同步性	数据完整性原则	导航	有用性
可用性	数据规模	效用性	时效性和可用性	有效率	

数据质量的维度

数据间错误或相互矛盾

一个或多个数据源的不同记录实际上标识同一实体。例如,企业的市场、销售和服务部门可能维护各自的数据库,这些数据会有大量的不同描述的重复客户。

在一年内可能过时。

数据
不一致

数据
精度低

数据
不完整

数据
陈旧

实体
不同一

- (1) **一致性**: 在数据集合中, 每个信息都不包含语义错误或相互矛盾的数据。
- (2) **精确性**: 数据集合中, 每个数据都能准确表述现实世界中的实体。
- (3) **完整性**: 数据集合中包含足够的数据来回答各种查询, 并支持各种计算。
- (4) **时效性**: 在信息集合汇总, 每个信息都与时俱进, 保证不过时。
- (5) **实体同一性**: 同一实体的标识在所有数据集合中必须相同而且数据必须一致。

不一致检测与修复

- 基于数据完整性约束
- 给定一组完整性约束 Σ ，发现数据库实例I中不满足 Σ 的部分，通过修复操作求解与I差距最小的I'，且I'满足 Σ
- 完整性约束主要用数据中各种依赖描述，例如包含依赖、函数依赖、条件函数依赖等

	CC	AC	phn	name	street	city	zip
t_1 :	44	131	1234567	Mike	Mayfield	NYC	EH4 8LE
t_2 :	44	131	3456789	Rick	Crichton	NYC	EH4 8LE
t_3 :	01	908	3456789	Joe	Mtn Ave	NYC	07974

$f_1: [CC, AC, phn] \rightarrow [street, city, zip]$, $f_2: [CC, AC] \rightarrow [city]$.

$cf d_1: ([CC = 44, zip] \rightarrow [street])$ \leftarrow

$cf d_2: ([CC = 44, AC = 131, phn] \rightarrow [street, city = 'EDI', zip])$ \leftarrow

$cf d_3: ([CC = 01, AC = 908, phn] \rightarrow [street, city = 'MH', zip])$ \leftarrow

- ✓ 修复 t_2 的街道和城市，使之和 t_1 相同
- ✓ 将 t_3 的city属性修复成"MH".

缺失值填充的方法

删除

直接删除相应的属性或样本。

统计填充

使用所有样本关于这一维的统计值对其进行填充，如平均数、中位数、众数、最大值、最小值等。

统一填充

将所有缺失值统一填充为自定义值，如“空”、“0”、“正无穷”、“负无穷”等。

平均数、中位数、众数、最大值、最小值

空、0、正无穷、负无穷

缺失值填充举例

年收入

在商品推荐场景下填充平均值，
借贷额度下填充最小值。

驾龄

没有填写这一项的用户可能是没有驾照，为它填充0较为合理。

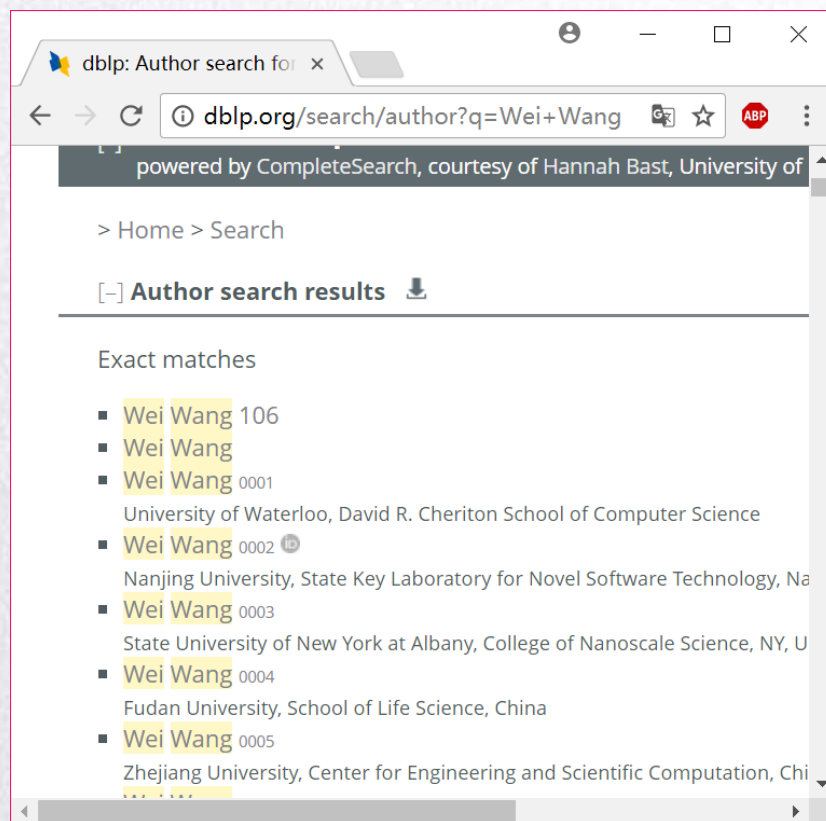
姓名	性别	电话	驾龄/年
王小明	男	18277777777	10
李刚	男	18266666666	

问题

当DBLP中检索“Wei Wang”的文章时，会检索到14个“Wei Wang”的197篇文章。

什么是实体识别

在给定的对象集合中，正确发现不同的实体对象，并将其聚类，使得每个经过实体识别后得到的对象簇在将现实世界中指代的是同一实体。



实体识别解决的问题

1. • 冗余问题 • 同一类实体可能由不同的名字指代。如名字叫王伟，用英文表示可能是“Wang Wei”，也可能是“ Wei Wang”。
2. • 重名问题 • 不同类的实体可能由相同的名字指代。例如在DBLP中检索“ Wei Wang”，会得到14个不同的作者。

Name	affiliation
Wei Wang	National University of Singapore
Wang Wei	National University of Singapore

Name	affiliation
Wei Wang	National University of Singapore
Wei Wang	Fudan University, School of Life Science, China

实体识别的两类技术

1. • 冗余发现 • 计算对象之间的相似性，并与阈值比较，从而判定对象是否属于同一实体类。
2. • 重名检测 • 利用聚类技术，通过考察实体属性间的关联程度判定相同名称的对象是否属于同一实体类。

Name	affiliation
Wei Wang	National University of Singapore
Wang Wei	National University of Singapore

Name	affiliation
Wei Wang	National University of Singapore
Wei Wang	Fudan University, School of Life Science, China

实体识别之后

经过实体识别之后，描述同一现实世界实体的不同元组被聚到了一起，然而这些对象的相同属性值可能包含冲突值。

真值发现

在这些冲突值中，发现真实的值。

Name	affiliation	Age
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	47

思路

投票方法

往往真值是由大多数的源提供。

Name	affiliation	Age
Wei Wang	Naonal Unrsity of Singpe	47
Wang Wei	Natiol Uniety of Sinaore	47
Wei Wang	National University of Singapore	41
Wang Wei	National University of Singapore	41

01

概述

02

大数据的来源与多源数据采集方式

03

数据集成和跨界应用的数据集成方法

04

数据的预处理

05

习题

习题

- 2.1 大数据的来源主要有几种？
- 2.2 大数据的集成的基本原理有哪些？
- 2.3 能否举例说明基于特征级别与基于语义的跨界数据集成方法的不同？
- 2.4 数据质量有几种维度？分别是什么？
- 2.5 你能提出一个金融行业领域中的数据获取的应用案例吗？
- 2.6 想实现对一个城市空气污染的检测和预测，请思考下述问题
 - (1) 需要哪些数据？
 - (2) 这些数据来源于何处？
 - (3) 这些数据应当以何种方式采集？
 - (4) 这些数据应当经过何种预处理？
 - (5) 如何集成这些数据以支持空气污染检测和预测的任务
- 2.7 请分析数据预处理应当在数据集成之前还是之后进行，为什么？
- 2.8 请分别举出在教育领域需要传统信息集成和跨界信息集成的实例。
- 2.9 请分析在交通大数据(如GPS采集的数据、打车软件中记录的数据)中可能遇到数据质量问题以及这些数据质量问题的检测方法和修复方法。
- 2.10 假设需要从大众点评、美团、百度外卖3个数据源收集北京市餐馆的信息，请简述可能会用到的数据集成步骤。针对上述场景，列举数据中可能存在的数据质量问题。