



西安电子科技大学
XIDIAN UNIVERSITY

第六章 数据集成

马 晶

经济与管理学院 信息管理系

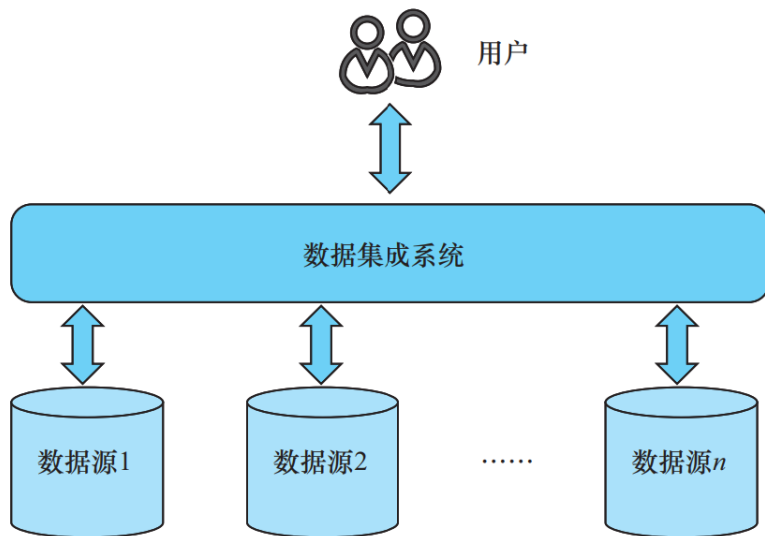
Email: majing@xidian.edu.cn

数据集成的定义

数据集成就是将若干个分散的数据源中的数据在逻辑或物理层面集成到统一的数据集合中。

具体来说，就是将不同来源、格式、性质的数据在逻辑或物理层面上有机地集成，通过一种一致的、精确的、可用的表示法，整合描述同一现实实体的不同数据，进而提供全面的数据共享，并经过数据分析、挖掘、产生有价值的信息。

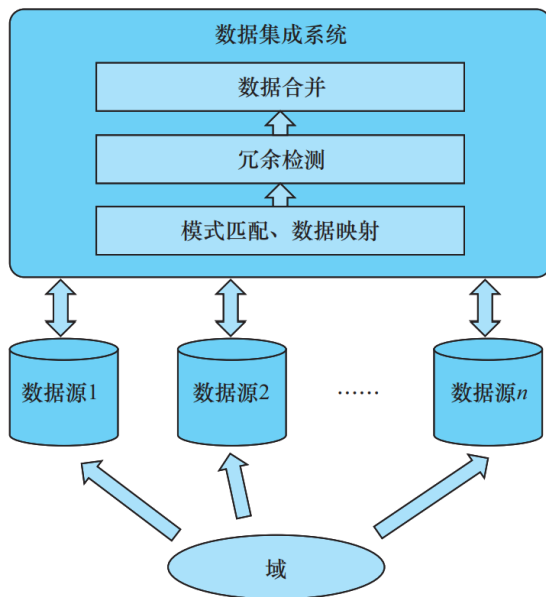
数据集成的核心任务是要将互相关联的分布式异构数据源整合到一起，使用户能够以透明的方式访问这些数据源。实现数据集成的系统称为**数据集成系统**，为用户提供统一的数据源访问接口，执行用户对数据源的访问请求。



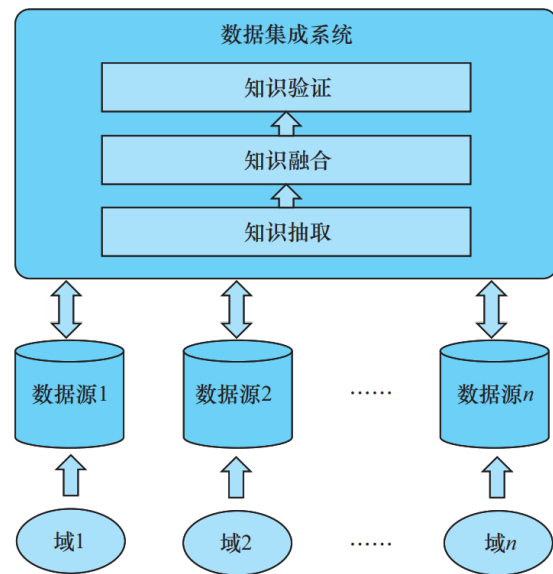
数据集成的分类

传统数据集成通过模式匹配、数据映射、冗余检测、数据合并等技术，通过统一模式访问将多个数据源集成起来。

除模式映射之外，还涉及多种知识融合的方法，面向这类需求的数据集成称为**跨界数据集成**。



(a) 传统数据集成



(b) 跨界数据集成

数据集成的难点

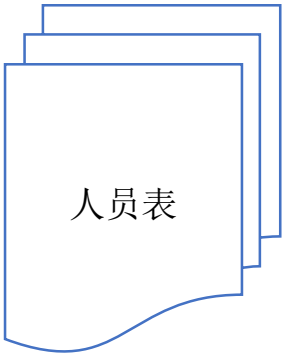
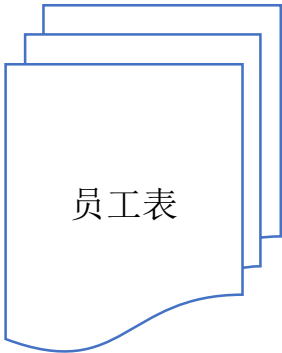
异构性

被集成的数据源通常是独立开发的，各数据源固有的异构性给集成带来很大困难。这些异构性主要表现在**语法异构**和**语义异构**上。

- ◆ **语法异构**：一般指源数据和目的数据之间命名规则及数据类型上的异构性。如果数据源是关系型数据库，那么命名规则通常指模式名（表名）和属性名（字段名）。语法异构**解决起来相对简单**，无需关心每条记录的内容和含义，对数据模式有了解，能够找到表、属性的恰当映射，就可以把不同表中的数据统一起来。
- ◆ **语义异构**：很难单纯从分析数据结构入手，其涉及对数据内容和含义的理解。语义异构的**处理难度要比语法异构大得多**，需要对数据含义做进一步分析，而且经常需要额外的领域知识作为辅助。



语法异构



	姓名
--	----

	姓		名
--	---	--	---

	出生日期
	1999-10-05

	出生日期
	1999年10月5日



语义异构

	张三		男		1998-4-23		汉		未婚

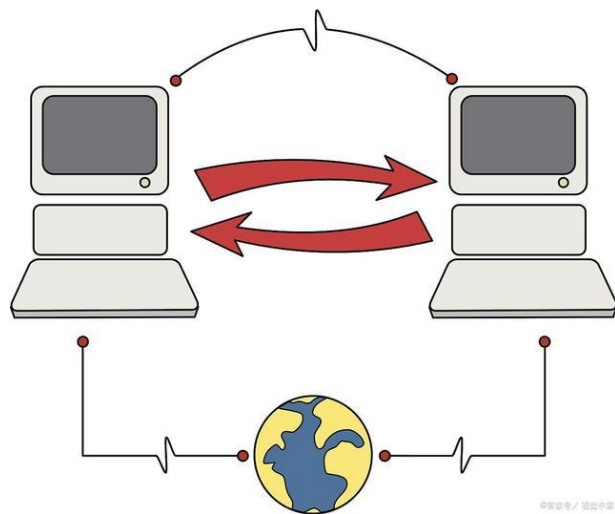
	Zhang San		男		1998-4-23		汉		Married

数据集成的难点

分布性

数据源往往是异地分布的，数据集成过程依赖于网络传输，如何保证传输过程的性能和安全性也是集成过程的难点。

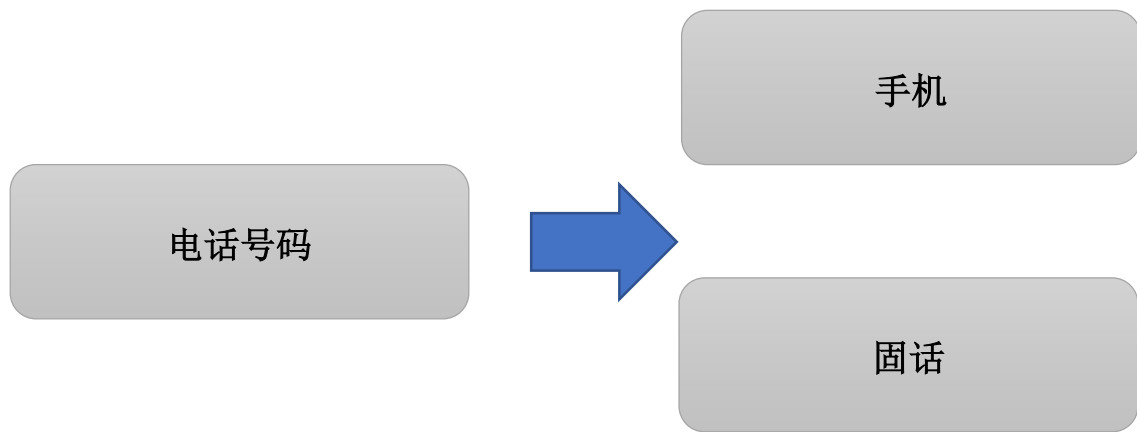
- ◆ 能不能传？出于某些安全性的考虑，并非所有的数据都适合通过网络进行传输。
- ◆ 传哪些数据？并非所有数据都需要参与数据集成过程。
- ◆ 传输速度能否满足要求？如果传输的数据量太大，还需要考虑传输速度是否能够满足要求。在数据量大的情况下，可能有一些加速的方法。



数据集成的难点

自治性

在一些应用中，不同的数据源有很强的自治性，数据源的拥有者有增加、修改、删除数据的权利，可以直接在本地修改数据而不通知数据集成系统。如果修改影响到了某些关键记录，或修改了关键数据模式的定义，那么就有可能导致数据集成出错。



传统数据集成

传统数据集成的主要目的是数据的共享。

传统数据集成的最重要的一点就是将来自源系统的数据统一匹配到目标模式。

形式化地，令 I 表示数据集成系统，则 $I = \langle G, S, M \rangle$ ，其中 G 是全局模式， S 是数据源模式， M 是全局模式和数据源模式之间的映射。

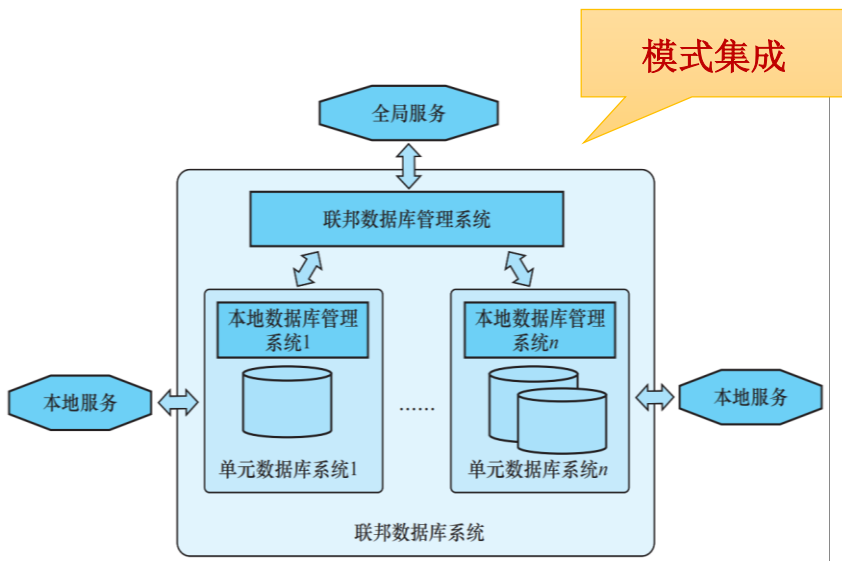
三种非常重要的传统数据集成方法：**联邦数据库系统**、**中间件集成**和**数据仓库**。

联邦数据库系统（federated database system, FDBS）是多数据库系统中的一种。主要目标是以透明的方式将多个自治数据库系统映射到单个联邦数据库中。构成联邦数据库的自治数据库系统称为**单元数据库系统**（component database system, CDBS）。单元数据库通过计算机网络互连，可以在地理上分散。联邦数据库系统将单元数据库系统按不同程度进行集成，从整体上提供控制和协同操作。

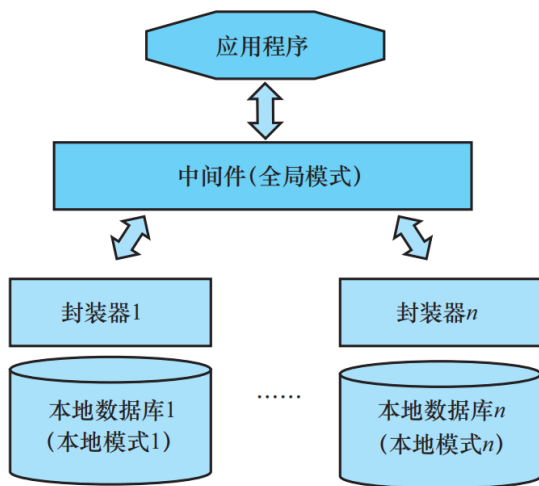
模式集成

联邦数据库分为**松耦合联邦数据库**和**紧耦合联邦数据库**。**松耦合**要求用户创建和维护联邦，联邦系统不做硬性限制。**紧耦合**联邦数据库由联邦系统提供统一的全局模式和数据访问接口，由全局数据库管理员创建并维护该模式。

特点：数据源独立存在，一个数据源可以访问其他数据源提供的信息。中间很多设计细节难以实现，需要慎重考虑。



中间件 (Mediator) 在软件架构设计中非常常见。中间件位于数据源系统（异构的数据库、遗留系统、Web资源等）和应用程序（服务调用接口、终端用户等）之间，向下协调各数据源系统，向上为应用层提供统一的数据模式和数据访问的通用接口。



中间件集成同样使用全局数据模式，通过在中间层提供一个统一的全局模式来隐藏底层数据细节，使得从用户的角度来看，数据层的数据源整个构成了统一整体。

中间件数据集成系统的核心在于中间件和封装器的设计。

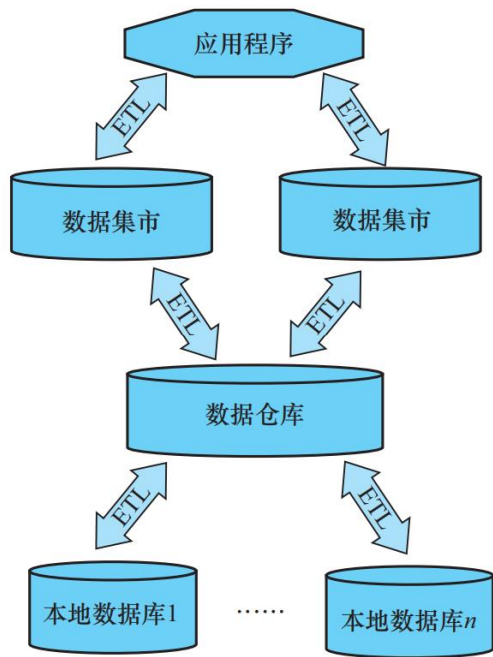
特点：注重全局查询的处理和优化，相对于联邦数据库系统的优势在于能够集成非数据库形式的数据源，有很好的查询性能，自治性强，但很多中间件集成是只读的，而不像联邦数据库系统那样对读/写都支持。

数据仓库 (Data Warehouse) 是用于生成报表和数据分析的系统，被认为是商业智能 (Business Intelligence) 的核心组成部分。数据仓库是一个中央数据存储库，集成了来自一个或多个不同来源的数据。通过将当前和历史数据存储在同一地方，可以很容易地为整个企业创建分析报告。

数据在最终存储之前，可能需要一些预先的处理。**提取、转换、装载 (Extract-Transform-Load, ETL)** 是将数据从一个或多个源转移存储到目标系统的通用过程。

- ◆ **提取**：旨在将数据转换为适合处理的单一格式，并验证提取结果的正确性。
- ◆ **转换**：将一系列规则或函数应用于提取得到的数据，以准备将其加载到最终目标系统中。一个重要功能是数据清洗，其目的是仅将“适当的”数据传递给目标。
- ◆ **装载**：将数据装载到最终目标系统中，原则上，该最终目标系统可以是任何数据存储系统，数据仓库只是目标系统的一种。由于需求不同，此过程的差异可能很大。

ETL几乎贯穿了整个数据仓库构建和使用过程。**数据集市**（Data Mart），有时也称为数据市场，包含为了满足特定需求存储的多维数据。数据集市中的数据是从数据仓库中抽取出来的一个子集，目标是为了某些（通常是重要的）应用提供高效、便捷的服务。



除了ETL，数据仓库的构建可能还涉及到ELT或CDC。

ELT是抽取、装载、转换（Extract- Load -Transform）的缩写，适用于数据结构化比较好、集成工作量不大的情况，主要思想是数据抽取后不立刻转换，而是装入暂存区进行清洗，最后在数据仓库中完成转换。

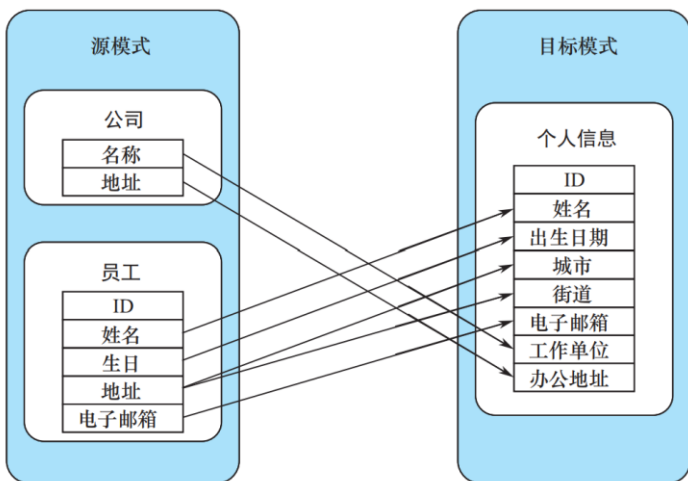
CDC是变化数据抓取（Change Data Capture）的缩写，通常通过在源系统上安装第三方应用程序来采集源系统的数据变化，多用于实时性要求非常高的场景。

传统数据集成的关键技术

模式匹配

在数据集成中，模式匹配的主要目标是完成不同数据模式之间的语义映射。这个过程需要两个主要步骤：

- (1) 需要识别数据对象的语义相关性，如包含关系或等价关系；
- (2) 需要完成语义相关的对象之间的映射，这些映射可能是组合的，并不一定是一一映射。



模式匹配的困难主要来源于数据的多源异构性：

- (1) 不同的模式可能使用不同的表示来指代相同的信息；
- (2) 使用同一个表示来指代不同的信息；
- (3) 对同一个信息的描述精度不一致。

模式匹配的方法

模式匹配的方法可以大致分为**模式级的匹配**、**实例级的匹配**和**混合匹配**，都是利用模式信息或模式和实例级别信息的方法。其中模式级的匹配仅考虑模式信息，而实例级的匹配还考虑具体的数据实例。

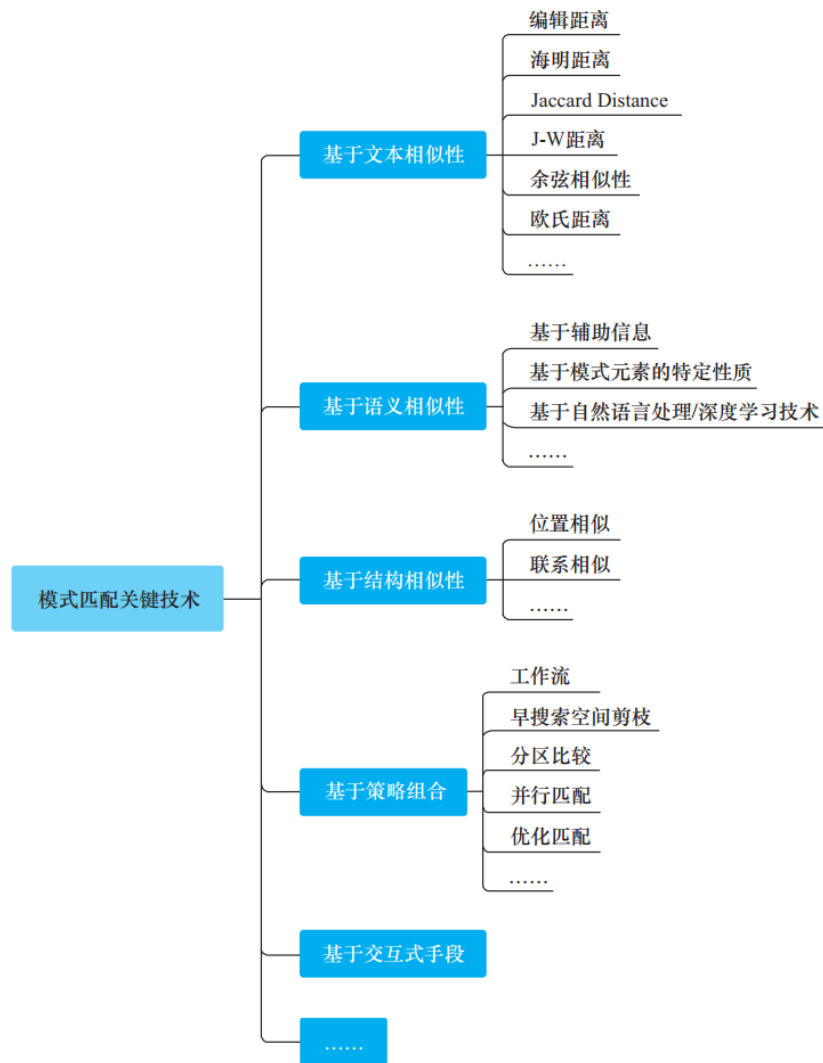
- ◆ **模式级的匹配**：考虑抽象的模式元信息，包括列名、元素名等，如名称、描述、数据类型、关系类型、约束和模式结构。具有相似属性的模式元素通常被看作匹配的元素。
- ◆ **实例级的匹配**：通过观察数据实例来确认模式元素的含义，在模式级可用的信息不足时，实例级的匹配可以增强对模式级匹配结果的置信度。
- ◆ **混合匹配**：直接结合多种匹配方法，将几个模式匹配的方法加以合并来基于多准则或信息源确定候选匹配，有些额外的信息可以用在混合匹配中。

模式匹配的关键技术

可以将自动化的模式匹配技术分成几个特定阶段，包括经典的模式自动匹配、复杂的模式自动匹配、针对特定领域的模式自动匹配及使用半自动的模式匹配工具等。

这些模式匹配技术需要理解模式的不同侧面，包括文本相似性、语义相似性、结构相似性等。

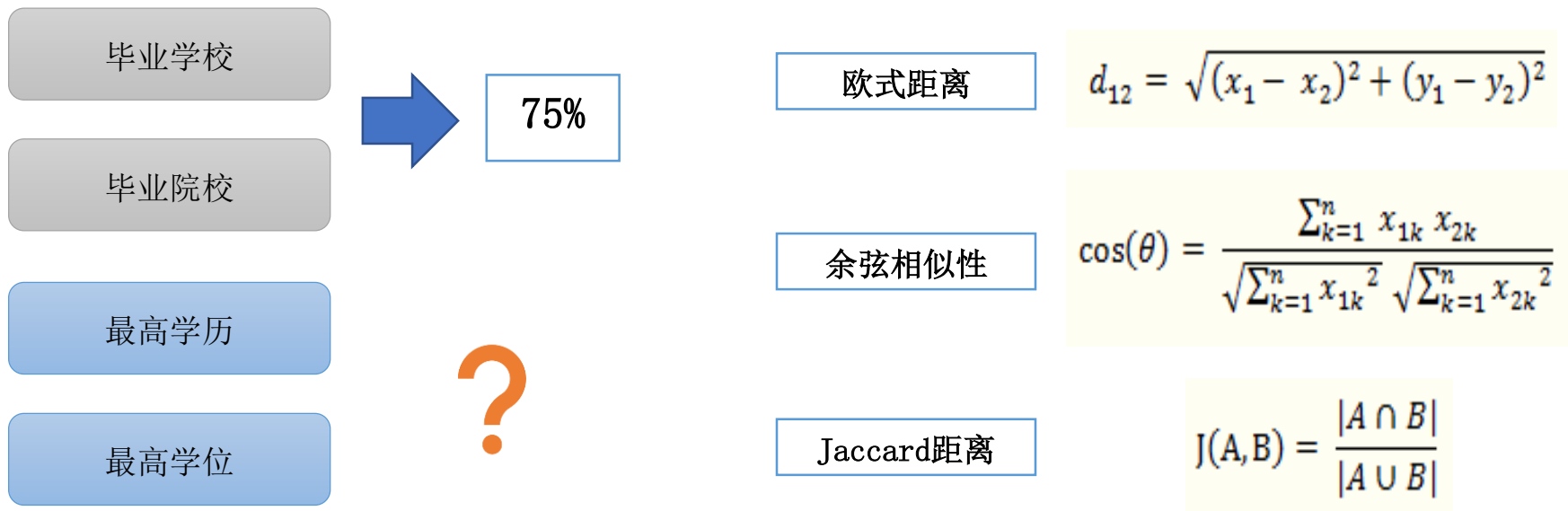
经典的模式自动匹配方法通常将目光放在某个特定侧面上。



◆ **基于文本相似性**的匹配通常基于名称、描述、分隔符、字符串、子字符串进行匹配。这种技术主要考虑文本上的相似性。基于文本相似性的匹配技术主要依赖于底层的文本相似度量。常用的字符串相似度量方法都可以用在这里。

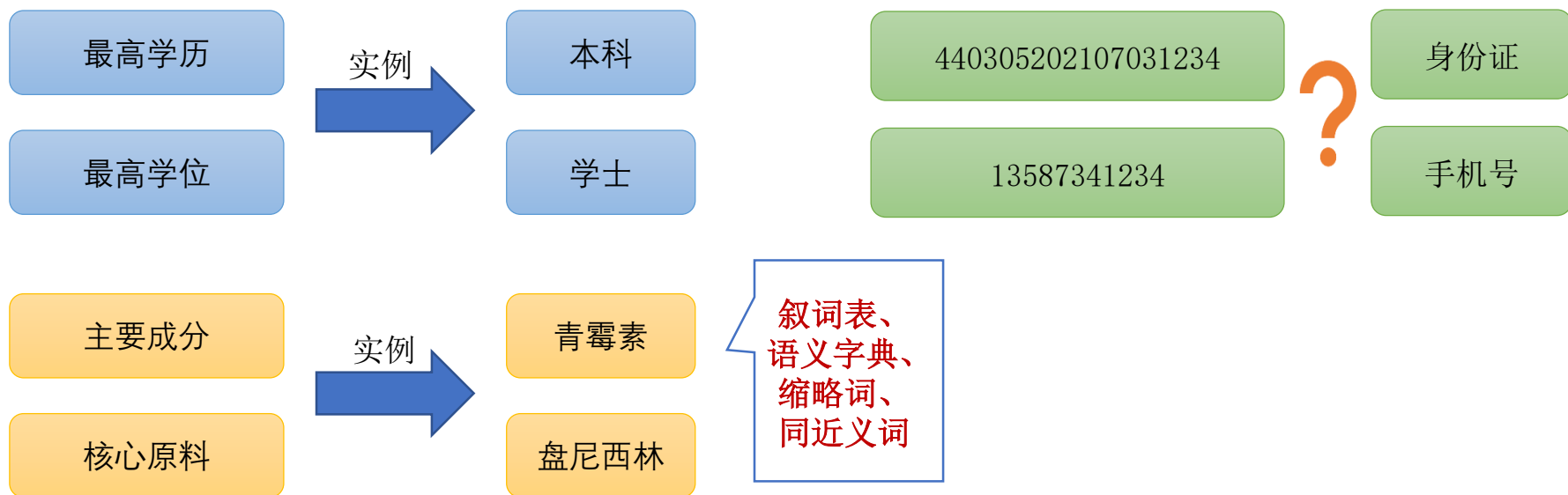
优点： 相对简单，易于实现和理解；由于只需要将模式元素当成字符串来比较文本层面的相似程度，因此自动化程度也较高。

缺点： 由于其简单性，很多涉及数据语义层面的匹配可能会出现遗漏或错误。



◆ **基于语义相似性**的匹配通常要利用一些自动化的语义理解技术来比较两个模式实例内容的语义相似性，从而推断两个模式元素的相似性。

可以通过一些辅助信息来帮助计算语义相似度。此外，也可以通过模式元素的特定性质来帮助判断语义是否一致。然而，辅助信息的获取、维护、更新往往很难自动完成。一些自然语言处理技术或许可以解决辅助信息依赖的问题。



◆ **基于结构相似性**的匹配通过分析**模式元素的结构特征**来判断其是否相似或匹配。如果两个模式元素出现在相似的组里（位置相似）或者有相同的联系、与相似的元素有关（联系相似）就可以认为这两个模式元素是相似的。

基于结构相似性的匹配方法**也需要额外的辅助信息**，用来指明哪些元素是相关的。

◆ **基于策略组合**的匹配将目光专注于模式的某个侧面通常有助于设计出小而精的算法，但往往会忽视模式其他侧面提供的对于匹配有利的信息。因此，可以使用一些策略来组合不同技术以期得到更优的匹配结果。

现有的组合策略有 **workflow、早搜索空间剪枝、分区比较、并行匹配、优化匹配**等。由于完全自动化的模式匹配很难保证匹配的准确度，因此基于交互式手段将用户的意见引入模式匹配任务也是较为常见的解决思路。

数据映射

在数据集成中，数据映射的目标是在给定两个数据模型之间建立起数据元素的对应关系。

数据映射在数据集成中，不仅用于完成数据源和目标之间的数据转换，还用于识别数据关系、发现隐藏的敏感数据和冗余数据等。

- 从源模式数据表的数据表示到目标模式数据表的数据表示的转换。
- 源模式数据表中的数据对于目标模式数据来说可能冗余。
- 源模式数据表中的数据可能需要进一步提取后才能集成到目标模式数据表中。

数据映射

对于不同的应用场景，映射任务的复杂度有所不同，具体取决于要映射的数据的层次结构及源数据库和目标数据库的结构之间的差异。为了完成数据映射，可能的思路大致有两类：

◆ **字符（串）层面映射：**通过观察数据在字符上的相似度来判断映射关系是否存在。这种匹配相对比较容易完成，自动化程度较高，因此在实际往往首先会从字符串层面考虑是否匹配。

◆ **语义层面映射：**需要考虑数据背后所表达的语义是否是一致的。为了完成语义层面的映射，需要判定不同的概念或实例的等价性。常见的等价性有：**类别等价、属性等价、实例等价**。



跨界数据集成

- 传统数据集成中，基本假设数据来自同一个域。
- 跨界数据集成要处理的任务更难，要对不同领域相关联的数据进行集成，基于不同领域产生的多个数据集中数据对象的隐含关联性融合数据，协同发现新知识。
- 跨界数据集成的难度在于必须理解来自不同领域的多模态数据。
- 跨界数据集成可以被分为三类，分别是**基于阶段的集成**、**基于特征的集成**和**基于语义的集成**。

基于阶段的集成

基于阶段的集成在数据分析、挖掘的不同阶段使用不同的数据集合。由于数据集是异步使用的，所以用在不同阶段的不同数据集可以是低耦合的，并不强求数据形式必须一致。

交通异常检测

导航数据 汽车行为
社交媒体 事故异常



基于特征的集成

基于特征的集成方法是指基于表征学习等方法，从不同数据集中提取出来的原始特征中学习出新的特征，把这种新的特征应用于分类、预测等数据分析挖掘任务。

这种方法的实现可以有直接关联、深度神经网络（DNN）等方法。

基于语义的集成

基于语义的集成方法需要清晰地理解每个数据集语义，要知道每个数据集代表着什么、为什么不同的数据集可以融合、它们之间怎样相互增强特征。

基于语义的数据集成方法可以分为基于多视图的方法、基于相似性的方法、基于概率依赖的方法和基于迁移学习的方法等。

智能化时代的数据集成

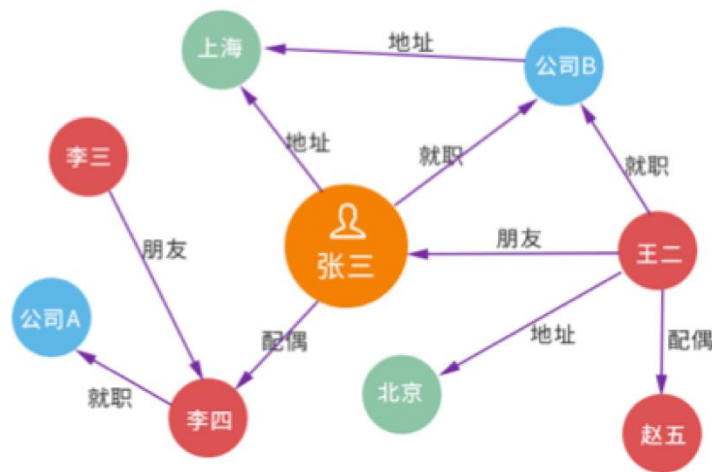
- 大数据时代，数据的丰富极大地促进了各类智能化方法的发展。一方面，人们希望从海量数据中提取信息、提炼知识，并不断扩大知识的边界；另一方面，人们也希望能够在扩大数据量的同时为数据提供必要的保护。

知识图谱融合

知识图谱是由**实体和关系**组成的多关系图，通过图结构来描述实体具有的各类属性和实体之间的多种联系，是一种结构化的知识表示方法。知识图谱通过规范化语义来汇聚知识，进而支持复杂关系数据的挖掘分析和推理。

目的：为人工智能技术赋能，让机器学习到知识的更多含义，从而具备一定的认知能力。

应用：大数据决策分析、搜索引擎、推荐系统和智能问答等众多领域。



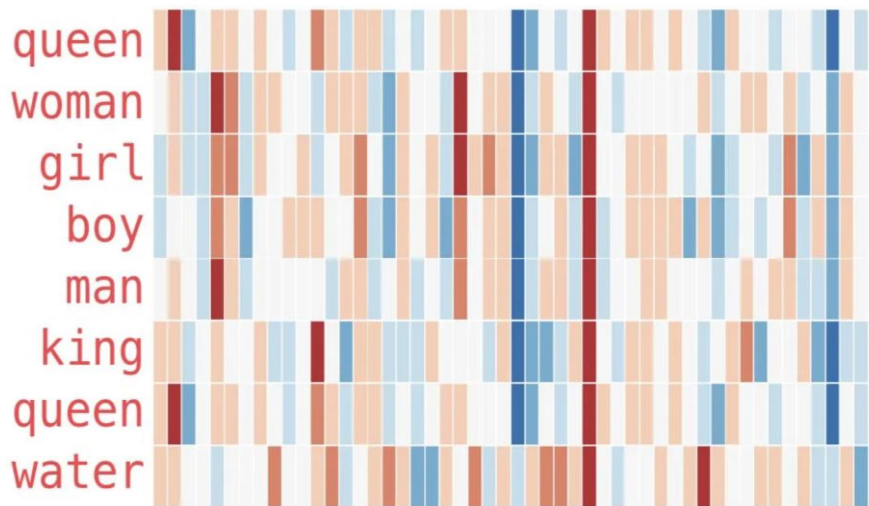
集成异构知识图谱的两种思路

- 先集成用于构建图谱的底层数据，将不同图谱的底层数据集成为一个大数据集合之后，再在新的数据集合上重新构建更完善的图谱；
- 不考虑底层原始数据，直接将多个异构图谱集成为一个更大、更完善的知识图谱。

对于第二种思路，如何将不同知识图谱中语义相同的知识关联并融合起来，就成了一个核心问题，即**实体对齐**。

我们可以将知识图谱形式化表示为 $G=(E, R, T)$ ，其中， E 为实体， R 为关系， T 为(头实体，关系，尾实体)三元组，对于两个不同的知识图谱，实体对齐的任务是要在两个知识图谱中找到并统一实际含义相同的实体。

语义平移现象



1.所有这些不同的单词都有一条直的红色列。它们在这个维度上是相似的（虽然我们不知道每个维度是什么）

2.你可以看到“woman”和“girl”在很多地方是相似的，“man”和“boy”也是一样

3.“boy”和“girl”也有彼此相似的地方，但这些地方却与“woman”或“man”不同。这些是否可以总结出一个模糊的“youth”概念？可能吧。

4.除了最后一个单词，所有单词都是代表人。添加一个对象“water”来显示类别之间的差异。你可以看到蓝色列一直向下并在“water”的词嵌入之前停下了。

5.“king”和“queen”彼此之间相似，但它们与其它单词都不同。这些是否可以总结出一个模糊的“royalty”概念？

king - man + woman \sim queen



?

首都?

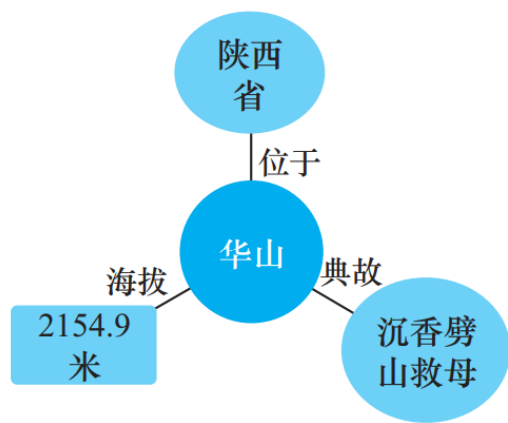
$$v(\text{中国}) - v(\text{北京}) \cong v(\text{日本}) - v(\text{东京})$$



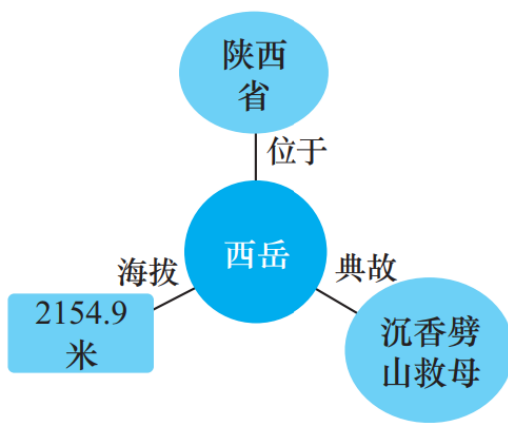
$$\begin{aligned} v(\text{中国}) + v(\text{首都}) &= v(\text{北京}) \\ v(\text{China}) + v(\text{Capital}) &= v(\text{Peking}) \end{aligned}$$

图神经网络

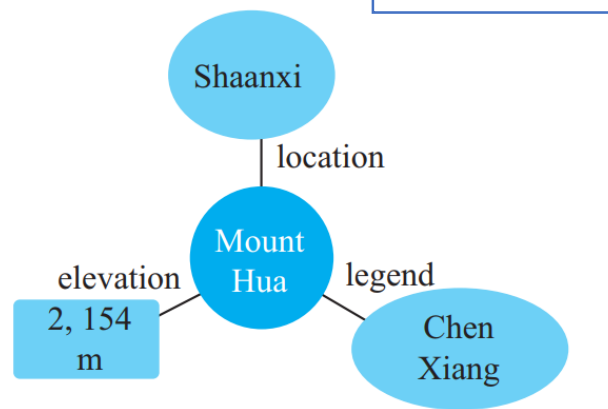
由于知识图谱本身可以视为图数据，因此一些处理图数据的前沿方法也被尝试用来处理知识图谱中的一些问题，例如图神经网络（Graph Neural Network，GNN）。核心思想是通过聚合图结构中的邻居信息来生成节点的向量表示。



(a)



(b)

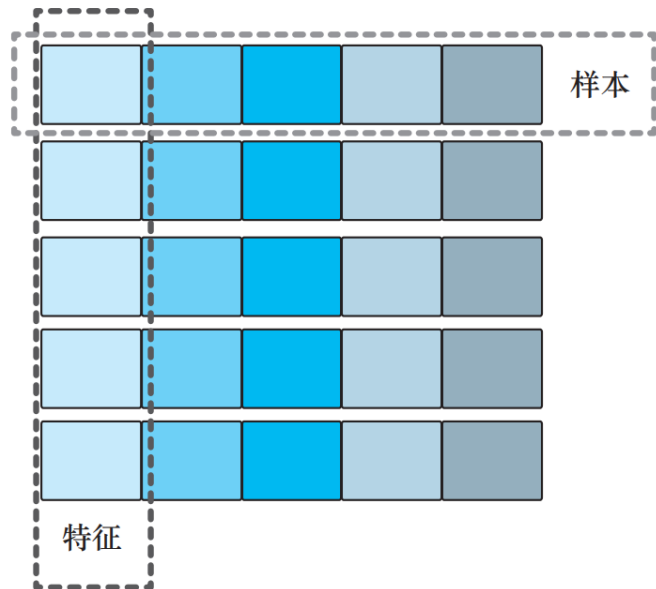


(c)

引入词典
或规则

联邦学习

- **联邦学习**（Federated Learning）的概念于 2016 年由谷歌提出，其尝试不集成“源头”而直接集成“结果”。具体地，联邦学习不直接集成数据，而是尝试基于加密机制，让参与方能够安全地交换模型训练的某些中间结果，不断迭代，最终完成在整个大数据集上的模型训练。
- 联邦学习保证任何一个参与方都不能基于公有模型反推出其他参与方的特征，即在保证训练的过程中各参与方都不暴露原始数据的基础上，基于各参与方提供的数据训练出比单一数据集上更强大的模型。
- 由于各参与方通常是业务关联的，因此其数据集之间可能有重叠的样本或重叠的特征。





本章要点

- 掌握数据集成的概念、分类，掌握传统数据集成的方法与关键技术
- 了解跨界数据集成
- 了解数据集成的新发展