【1】

欢迎同学们学习大数据预处理，现在开始讲解第一章：大数据预处理总述。

【2】

在开章之前，我们首先来看下，在一个数据挖掘项目中，有哪些主要的开发流程呢？如图所示，我们可以看出，数据预处理是在数据采集之后，数据存储和分析挖掘之前的。数据预处理的工作就是读入数据采集阶段的数据，然后采用各种手段清洗脏数据，再输出满足数据分析、数据挖掘的数据集。这个数据集一般都是存储在分布式系统中。当然，对这个数据集进行分析挖掘后，得出的分析挖掘成果，是通过可视化等手段向客户展示的。接下来，我们进入到预处理第一章。

【3】

这一章的脉络是这样的，首先阐述下数据预处理的背景，也就是在数据挖掘项目中，为什么要进行数据预处理，那就是因为数据的质量存在各种问题，比如缺失，重复等等。所以，我们就明白了数据预处理的目的就是提升数据的质量，把数据中的脏数据全部清除掉。通常，我们面对的数据源是多样化的，怎么去清除脏数据呢？在业界，通常是通过数据清理、数据集成、数据归约、数据离散化等一系列的流程步骤完成的，最后，我们将介绍，在数据预处理的工作中，我们可以使用什么工具进行数据清理。这个工具就是Kettle和Python语言，它们是本书的两大核心内容。

【4】

我们首先看下第一节，数据预处理的背景：那就是存在各种的数据质量问题。到底存在哪些具体的问题呢？

【5】

数据的质量问题主要存在这六大方面的问题：准确性、完整性、一致性、相关性、时效性和可信性可解释性

这6个方面的问题，可以从不同的角度去划分。

如果从采集的角度划分，可分为准确性、完整性、一致性。

如果从应用的角度划分，可分为相关性、时效性。

如果从用户使用的角度划分，可分为可信性和可解释性。

【6】

【6.1】现在，我们从采集角度来看的第一个问题：准确性.

准确性是指数据是否正确的，数据存储在数据库中的值是否对应于真实世界的值。例如，某用户在使用支付宝绑定银行卡时，网站要求验证用户的真实姓名和身份证号码。如果用户提供的证件号码与实际号码一致，那么该号码存储在数据库中的值就是正确的。

【6.2】数据的不准确由哪些**原因**造成呢：

一是在收集数据时，**设备出现故障**，导致数据存储的值出现乱码；

第二是在**数据输入**时，人为的输入不准确的信息，或者计算机内部出错导致录入的信息有误，比如我们上网注册一些信息时，出于隐私考虑，用户会故意输入不正确的信息，包括年龄、地址、手机号等。

第三种是在**数据传输**的过程中出现错误。比如，超出了传输缓冲区的大小，数据会出现截断等现象。

最后一种是命名约定、数据代码、输入字段的格式不一致导致出错。其中，最常见的是：不按格式输入导致出错，例如输入字段为日期时，多个用户输入日期的格式不一致。

这些都是数据不准确的原因。

【7】

【7.1】第二个数据质量的问题是数据完整性。完整性是指信息具有一个实体描述的所有必需的部分。在传统关系型数据库中，完整性通常与空值（NULL）有关。空值是指缺失或不知道具体的值，可能是一条记录中的某个属性缺失，也可能是整条记录都丢失。

【7.2】导致数据出空缺值得原因有：

（1）因涉及个人隐私，无法获取相关属性的数据。

（2）有些信息虽然不涉及隐私，但在采集时被认为是不必要的，所以当时没有记录，这就是人为的疏漏。

（3）数据输入或传输时，由于机器的故障导致。

不完整的数据对数据分析会产生影响，比如考虑构造一个预测交通事故发生率的模型。如果忽略了驾驶员的年龄和性别信息，那么除非这些信息可以间接地通过其他属性得到，否则模型的精度可能是有限的。这种情况下，我们就需要尽量采集全面的数据信息。

【8】

第三个数据质量的问题是一致性。数据一致性是指在数据库中，不同表中存储和使用的同一数据应当是等价的，表示数据有相等的值和相同的含义。

比如表1描述学生的基本信息，包括学号、姓名、性别、年龄和所在专业，而所在专业必须从专业信息表获取。表2描述了专业的基本信息。从这两个表可以看到，表1中的学生王小刚所在的专业号并没有出现在表2中，说明该条记录的专业号有误，必须修改正确，才能保证两张表对应字段的正确性，这是数据的逻辑不一致。

【9】

数据不一致还体现在记录的不规范上，比如两个表中对日期的格式记录不一致，这导致在数据集成中造成数据冲突。

如图所示，日期的格式有很多种，在数据预处理时需要统一格式。

另外还有不同数据库对同一属性的命名规则不一样。

比如如性别：在数据库A中 male=1 , female=2 ，而数据库B中为 male=‘男’ ，female=‘女’，数据库C中为 male=‘M’ , female=‘F’。

【10】

另外在数据出现冗余的情况下，数据内容由于各种原因比如**并发控制不当**，或**程序故障导致前后数据**不一样也是造成数据不一致的原因。

还是以学生信息为例，**学号95004**的学生所在专业号无法在专业信息表中找到，可能是由于**专业信息表之前是存在M15这个专业号的，但是由于某种原因将专业号专业号M15修改成了M05**，而系统因为并发控制不当或程序故障导致**没有对学生信息表中的专业号进行同步修改**，这就照成了 数据前后不一致。

【11】

现在，我们从数据的应用角度看数据存在哪些质量问题。

第一个是相关性的问题，数据的相关性是指数据与特定的应用和领域有关。

与数据相关的应用场景一般有，比如进行数据挖掘或构造模型预测时，需要采集相关的数据。例如考虑构造一个模型，预测交通事故发生率。如果忽略了驾驶员的年龄和性别信息，那么除非这些信息可以间接地通过其他属性得到，否则模型的精度可能是有限的。这种情况下，我们就需要尽量采集全面的相关的数据信息。

另外一个相关性的质量问题表现在相同的数据，在不同的应用领域中，相关性也是不一样的。例如，对于某个公司的大型客户数据库，由于时间和统计的原因，顾客地址列表的正确性为80%，其他地址可能过时或不正确。当市场分析人员访问公司的数据库，获取顾客地址列表时，基于目标市场营销考虑，市场分析人员对于该数据库的准确性满意度较高。而当销售经理访问该数据库时，由于地址的缺失和过时，对该数据库的满意度较低。

【12】

数据的时效性是指有些数据会随时间而变化的，

这些数据收集后就开始老化，使用老化后的数据进行数据分析、数据挖掘，将会产生不同的分析结果。

常见的场景有商品推荐。顾客的当时购买行为或Web浏览行为称为快照，它只代表有限时间内的真实情况。如果数据已经过时，则基于它的模型和模式也就已经过时，所以进行商品推荐需要采集当前的数据进行分析和推荐。在这种情况下，我们需要考虑重新采集数据信息，及时对数据进行更新。

另外一个典型的应用场景是城市的智能交通管理。以前没有智能手机和智能汽车，很多大城市虽然有交管中心，但它们收集的路况信息最快也要滞后20分钟。用户看到的，可能已经是半小时前的路况了，那这样的信息就没啥价值。但是，能定位的智能手机普及以后可就不一样了。大部分用户开放了实时位置信息，做地图服务的公司，就能实时得到人员流动信息，并且根据流动速度和所在位置，区分步行的人群和汽车，然后提供实时的交通路况信息，给用户带来便利。这就是大数据的时效性带来的好处。

【13】

【13.1】数据的可信性由三个因素决定：数据来源的权威性、数据的规范性、数据产生的时间。例如新浪微博某一用户发布的微博内容是否具有可信性，首先确定数据来源是否具有权威性，如果是权威机构的数据，那么可信度比较高。如果微博字数较长且叙述比较详细，可信度也会增加。同时微博的发布时间是否接近实时，也影响数据的可信度。

【13.2】最后数据的可解释性，也称为可读性，是指数据被人理解的难易程度，如果数据具有解释性或包含有注释性信息，而且数据书写规范，则数据的可解释性越高。相反如果数据晦涩难懂就根本不具备分析的条件。

【14】

接下来，我们到了第二节，数据预处理的目的：正因为数据存在各种质量问题，所以我们需要提升数据的质量。

【15】

【15.1】数据预处理的目的是由它在数据开发项目流程中所处的环节决定的，这个环节是比较重要的，不可或缺的，它是数据分析挖掘的输入。

【15.2】所以总体来说，它的目的是提升数据的质量，满足数据分析挖掘的输入。

分开来讲，有三点。一是改进数据的质量，提高数据挖掘过程的准确率和效率，二是保证数据挖掘的正确性和有效性。这两点都是由于脏数据的影响造成的。三是通过对数据格式和内容的调整，使得数据更符合挖掘的需要。也就是通过数据变换、数据归约等手段，使得数据更符合数据挖掘算法的要求。

【16】

本节到此结束，本节视频先阐述下数据预处理的背景，也就是在数据挖掘项目中，为什么要进行数据预处理，那就是因为数据的质量存在各种问题，这些问题可从采集角度，应用角度，用户角度去区分，分别为准确性、完整性、一致性、相关性、时效性、可信性和可解释性。因为这些问题的存在，影响了数据分析、数据挖掘的效果，所以，在数据预处理的目的就是提升数据质量，符合数据挖掘的需求，保证数据挖掘的正确性和有效性