****

**什么是数据处理？**

数据处理有广义和狭义两种理解，广义的理解，所有的数据采集、存储、加工、分析、挖掘和展示等工作都可以叫做数据处理；而狭义的数据处理仅仅包括从存储的数据中通过提取，筛选出有用数据，对有用数据进行加工的过程，是为数据分析和挖掘的模型所做的数据准备工作。

一般意义上讲的数据处理是狭义的定义，即对数据进行增、删、改、查的操作。在目前大数据的背景下，我们数据处理工作往往是通过技术手段来实现，比如说利用数据库的处理能力，对数据进行增加、删除、改动、查询等处理。

在实践中，数据处理工作中最大的是对数据进行清洗，即对不清洁的数据进行清洁化的工作，让数据更加规范，让数据的结构更加合理，让数据的含义更加明确，并让数据处在数学模型的可用状态。

**数据之“脏”**

数据的“脏”是一个比喻的说法。我们把数据记录不规范、格式错误、含义不明确等叫做数据的“脏”，包括几种典型的形式。

**（1）数据不规范的情况。**

比如姓名，同样是张三，有的地方记录为“张三”，有的地方记录为“张 三”，为了让两个字的姓名和三个字的姓名都有相同的长度，中间添加了空格。这种情况同样发生在地址字段里，比如说“北京”、“北京市”、“北 京”，虽然都是指北京，于我们人来讲很容易识别，但对计算机来讲，这三个写法代表着三个不同的值，我们需要通过建立映射关系的方式，将数据记录进行统一。

常见的数据不规范的情况还包括日期格式的问题。日期格式常见的几种记录方法如下：

2015/10/20

2015-10-20

2015年10月20日

10/20/2015

Oct. 20, 2015

October 20, 2015

2015.10.20

每个人都有不同的喜好和记录方法，这给计算机识别造成了很大的困难，一个公司的所有数据都应该有一个明确的规定，统一数据的录入格式。

**（2）数据不一致的情况**

数据不一致的情况往往是没有遵循我们单维数据表的原则导致的。因为同一条信息记录在不同的数据表甚至数据库中，当我们对数据信息进行更改之后，因为没有同时对所有的记录点都做相同的更改而导致的数据不一致的情况。为了避免这种情况，我们引入了“单维数据表”的概念，强调了公司内部同样一条信息，只能记录在一个地方，当其他地方需要的时候，通过索引查询的方式来保证数据的一致性，在任何数据表中存在其他表中数据来源的时候，都要在查询输出时进行“同步”更新。

数据的一致性虽然技术上比较容易实现，但在企业实践中却有着巨大的难度。采购部门会录入供应商的信息，而财务部需要向供应商付款，所以也会保留供应商相关的信息数据，而采购部门和财务部分分属不同的职能部门，财务部门也会采集一部分供应商财务相关信息，包括银行信息、账号信息、税务信息、工商信息等，如果发生变化，比如说法人变更、业务变更、企业性质变更等，财务会对其数据进行更新；采购部也会对供应商的信息进行采集并登录相关的信息管理系统。如果采购的管理信息系统能够同财务所使用的管理信息系统对接且能够把同条信息能够关联或者建立索引关系，则该公司的数据一致性比较容易保障。但如果两个都采用了不同的系统，就容易导致数据不一致的情况。而这种情况在大多数公司种都存在，且很严重。

**（3）标准不统一的问题**

我们对一些事物的描述方法需要建立统一的标准，从而让计算机可以有效地对文本数据进行处理。

举一个具体的例子。比如说导致产品出现质量问题的原因，多数情况下是手工录入的，同样的原因，不同的数据录入的描述会有不同。同样是因为电压不稳导致的产品质量问题，有人会录入为“电压不稳”，有人会录入为“电流不稳定”，有人录入为“供电问题”，有人录入为“缺少稳压设备”，有人录入为“供电负载异常”。。。如果没有统一的规范，我们在统计汇总时会产生上千个导致产品品质问题的原因。这给数据解读和分析，以及寻找改善措施带来很大的麻烦。

这就需要数据库管理员根据公司的实际情况，将该类原因进行归类，然后设定几个类别，由员工在系统中进行选择，而不是让他们手工录入。一般情况下，出现最多的前10名原因能够覆盖90%以上的情况，在录入中，先让员工选择，然后留出一个“其它”，当员工选择其它的时候再进行录入，这样就能够有效规范这种数据的录入标准化问题。

根据大多数人的记忆习惯，在经常使用的范围内，一般能够轻松记住7个左右的信息，因此，我们尽可能把这些导致质量的原因找出最常见的7个，对录入数据的人进行培训，他们基本能够记住这7个,特别是在面对教育水平不高的一线工人的时候。

**（4）格式不标准的问题**

所谓的格式不标准的问题是在数据录入时，使用了错误的格式。比如说，录入日期时，因为格式不规范，计算机不能自动识别为日期格式，出现了各种个样的文本；比如说录入数值时采用的中文字符格式，用了全角字符等，A和Ａ是不同的，1和１是不同的，0和O是不同的等；有些数据格式要求英文逗号分隔而错误地使用了中文的逗号；有些要求使用减号作为连接符，有的使用了下划线或者全角字符的连接符；有些要求使用英文引号，但录入时采用了中文的引号等等。

这种问题比较容易处理，需要信息系统设定相关的数据校验，如果录入不准确，数值录入为全角字符后会被识别为字符，系统弹出数据录入格式错误的警告基本可以解决大部分这样的问题。

**（5）附加字段的问题**

我们在数据清洗的时候，往往需要添加新的字段以便我们数学模型可以直接处理数据。比如说司龄、年龄等，我们数据库中可能没有直接的字段来记录员工的司龄，我们需要通过入职日期到目前数据采集日期间的差来计算司龄，这就需要添加司龄字段之后，通过入职日期来计算，年龄则通过出生日期来计算。

**数据杂质和噪音**

在外部大数据中因为数据价值密度较低，数据的杂质和噪音很多，需要大量的数据处理工作才能将有价值的数据和信息提炼出来，而企业大数据，特别是内部采集的数据，其价值密度高，几乎所有的数据和信息都是有价值的，其杂质和噪音也会少。

**什么是数据杂质呢？**

所谓的数据杂质就是在数据集中出现了与数据记录本身无关的数据，就如大米中出现了沙子一样，需要在处理数据的过程中，将这部分数据剔除。

比如说录音或者录像数据，本质上上为了记录经营或者管理活动，但在过程中可能因为没有活动发生，但录音和录像还在继续，这部分数据就会成为杂质。企业生产线上的监控录像，当没有生产时仍然在录像，拿这一部分时段的录像就可以从整体数据中剔除。就如行车记录仪，当停车时，记录仪检测到汽车已经不动超过10秒钟，录像就暂停，当图像中的画面有动时，则及时启动录像过程，这是一种比较智能的方式在遴选数据的采集和记录。

另外一种数据的来源是数据采集或者记录过程的杂质。比如说问卷调查，在问卷正式进行之前，编制问卷的人首先要做几遍测试，还会找其他人做个测试，以保证正式发布调研之后能够无差错，这部分的数据也会被调研系统后台记录，这些数据可以称作杂质，在处理调研数据集的时候，需要剔除。而调研的过程中，有人打开了调研链接，但做到一半就因为其他事情耽搁了，稍后又重新从头开始做该调研，则前面这部分未完成的问卷可以从数据集中作为杂质去除。

数据的杂质其实有很多种，具体数据采集的方式和方法不同，都会有不同类型的数据杂质进入到数据集，数据分析人员需要根据实际情况进行甄别。

**什么是数据噪音呢？**

所谓的数据噪音就是貌似与有用数据集，但仔细查看后并非该数据集该有的数据，或者仔细分析后没有价值的数据，当然也有一部分是我们无法解释其与其他数据差异的数据。与杂质不同，噪音是貌似相关的数据，但其实价值不大或者根本没有价值。

现在的电商是靠流量和销量说话的时代，特别是天猫和淘宝数据，购买者更加关心卖家的信用。卖家为了获得消费者更高的关注和购买量，往往采取“刷”信用的方式在提高自己的星级。对于电商来说，这些“刷”的交易数据，都可以看作是噪音数据，虽然这部分数据对于其它的分析可能非常有价值。

比如说一个订单数据集，在这个数据集中有一部分是内部测试形成的，也有是竞争对手测试形成的，还有可能是消费者测试网站形成的，有的甚至是数据采集机器人后台下单并取消，但是在这个过程中采集相关数据的，这一部分数据就可以看作数据的噪声，并非真正的交易数据。

公司在网上做了一次推广，短期内访问量大幅度上升，其中有部分访问量是竞争对手、品类爱好者、研究人员等进行的测试性或者信息获取性的访问，这部分访问就是我们研究客户访问及转化率的噪音。

**什么是数据清洗？**

所谓的数据清洗就是对原始数据进行规范化的处理，减少数据噪音，消除数据的不一致性，并对某些数据进行加工，以便数据处理软件和数据模型能够直接使用。数据清洗是数据处理工序之一，目的是提高数据的质量，为数据分析准备有效数据集。

数据清洗的方法有很多，主要与我们所使用的数据处理工具有关系。比如我们使用MS Excel,我们可以对数据进行查找替换、填充、分列、映射(vlookup)、透视等，如果规律性很强数据量很大的时候，我们还可以采用VBA编程的方式来实现。其它软件工具的数据清洗方法不一而足，需要熟练掌握对应软件的操作方法。

实践中，数据清洗工作是占用数据分析师时间最长的工作，虽然工作的价值产出很低，耗费大量时间，但这个工作必不可少，主要的原因是数据建表和数据采集过程中质量不高导致的。如果我们在数据采集、数据存储和数据传输过程中，提高数据的质量，保证数据的有效性，我们数据清洗工作可以大幅度缩减。而这个过程中，数据采集的方式、方法，以及自动化智能设备的使用是大幅度提高数据质量的关键手段。

要想在数据清洗上节省人工，需要数据系统中加入数据的校验，并制定相关的数据规范，让数据质量在源头录入的时候就是规范的，高质量的，即使是一些用户端口的数据，在录入的时候也要加入校验工作，通过示例的方式提醒用户按照一定的规则来录入。

我们经常见到一些网站在让用户录入姓名时要求用户录入姓和名，但是如果不进行校验，或者提示用户，用户很可能将姓氏录入到名字中，将名字录入到姓氏中，导致未来数据分析的时候存在问题。比如欧阳峰，如果峰字被录入到姓氏中，系统需要通过后台字典，提示用户——“您确信您姓峰？”，这种提示虽然消弱了用户体验，但对于数据的准确性还是非常有益的，语言上诚恳些，对待客户礼貌些，或者可以获得用户的理解。

**如何提高数据清洗速度？**

根据我们的实践经验，数据清洗工作占我们数据分析师工作量的70%，甚至以上，而且数据质量越差，这个比例越高。

其实提高数据清洗速度最有效的方法就是对数据采集和数据记录的规范性进行有效治理，从源头把控数据质量。如果数据源头的数据质量不高，数据清洗工作不仅会洗掉脏的数据，甚至还洗掉了某些有价值的数据，导致数据信息量的损失。

程序化方法是提高数据清洗工作效率的有效手段。我们往往面对的数据集比较大，如果手工一个个检查并清洗，可能需要耗费大量的人工时间。如果我们能够对数据不规范、不完整或者不相关的数据有较好的分析，总结他们中可能存在的规律性，然后用软件程序自动化完成数据的清洗工作，能够大幅度提升我们数据清洗的效率。寻找数据的规律性是用程序代替人工清洗的基础。

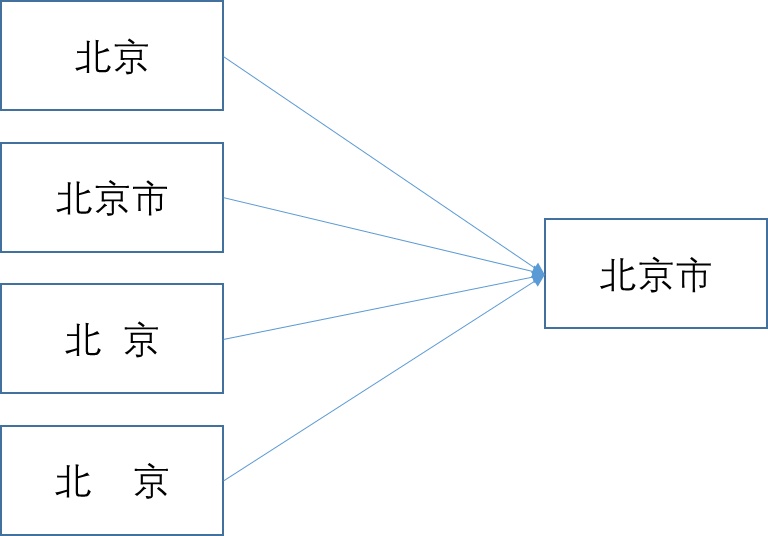
即使是使用Excel对数据进行清洗，如果能够用透视表+映射表的方式，会比手工查找+替换的方式要快很多。有的公司已经将一些常见的数据清洗方法编制成软件，但清洗的效果还是非常不理想，虽然这样的数据清洗软件能够大幅度节省人工的投入，解放数据分析师大量的工作。但这些软件一般都非常昂贵，一套软件在百万以上，能够快速处理数据，但仍然需要大量的人工干预。

数据清洗工作另外一个非常重要的原则就是：**永远给自己留下反悔空间。**

首先，尽量不要破坏原始数据。不能在原始数据集上直接改，如果修改丢掉了某些有价值信息，可能很难再找回来；如果发生了错误，将可能是灾难性的。所以：**先备份后清洗。**如果我们想规范日期格式，我们要在Excel中添加一列，让之前的日期列数据保留着，如果看着不舒服，可采取隐藏的方式，但直接删除或者替换都是不可以的。

其次，每次改变数据之前做好备份。我特别强调在对数据进行清洗时，禁止使用“查找+替换”的方式，因为这种方式改变了原始数据，如果发生错误，而Excel的Undo功能不能启用则麻烦就大了，即使保留了原始数据副本，可能之前的清洗工作会白费了。当数据量非常大的时候，任何有可能对数据集发生改变的操作之前都要做好备份工作。

映射表是一个非常好的操作方法，在利用Excel对数据进行清洗的时候，可以将同一字段的数据制作一个映射表，然后让Excel根据映射表对数据进行查找替换，我们常使用的功能是vlookup()函数。比如说地址中城市的名称，如果用户在填写的时候不是通过下拉表选择的，肯定会被填写的五花八门，人工能够识别，但机器不可以识别，所以可以通过透视表功能将所有的地址城市做个统计汇总，然后根据人工识别来建立映射表，然后再把原始的地址映射回去，从而将地址中城市名称标准化为一个唯一值，再对数据以城市为单位进行统计汇总时，数据才会准确。



利用第三方程序来进行数据的清洗也是一种方法，多数第三方数据清洗工具软件都是构建一个映射表，根据数据的特点进行猜测、精准匹配，并用后台“字典”来映射数据，然后将规范化的数据输出出来。第三方软件在适用性上往往都存在一定的缺陷或者说每个第三方程序都比较适合一类数据集，有的比较适合客户数据的清洗，有的比较适合产品订单数据的清洗，有的比较适合清洗社交媒体网站的数据。在选择第三方数据清洗软件的时候，要进行评比，用一个比较小的数据集进行测试之后再购买。这类第三方软件一般都比较昂贵，动辄几十万或者上百万，所以选择要慎重，并最好购买对方的服务，当数据清洗质量不佳的时候，让对方的技术人员参与，制定更加符合自己数据集的“字典”。目前国际上比较先进的该类工具软件在国内都不太好用，主要是中文的词语结构等问题导致不适应，大多数的数据清洗软件工具都是针对文本和数字类的数据的，中文的语种结构同西方的各种语种有着较大的差别，所以在数据清洗程度上有一定的局限性。通过购买软件公司的服务可以让数据清洗后的质量得到优化，如果数据量级达不到TB级别，购买服务要看值不值得了。

全文摘自《企业经营数据分析》赵兴峰著，转载已获得作者认可。

数据处理 数据清洗 数据采集 数据存储 数据分析