关于数据挖掘过程中数据清洗的研究

张祥飞
北京圆融科技有限公司　北京　100036

摘　要：数据挖掘简单来说，就是将所有的数据整合出来，找到并整合出来，因此，在学习模式识别，我们就要学习各类学科，例如，统计学、管理学、数据库等，因此，在当代社会数据挖掘技术也发展的越来越迅速，人们也越来越喜欢用挖掘技术和数据仓库技术来整合数据，一旦数据挖掘过程中发现这些数据有可以利用的价值，数据仓库技术就会将这些数据整合起来，数据清洗则是将错误的数据或是脏数据进行整理，因此在数据挖掘的过程中必须加上数据清理才能让数据库中的数据保证其真实性和有效性。因此，我国在发展数据挖掘过程中，还应该有很多学习和改善的内容，我国应该不断建立健全数据挖掘和数据清洗的各项策略研究。

关键词：数据挖掘；数据清洗；脏数据

Research on data cleaning in data mining

Zhang Xiangfei
Beijing Yuanrong Technology Co., LTD., Beijing 100036

Abstract: In simple terms, data mining is to integrate all of the data, to find and integrate, in learning, pattern recognition, therefore, we will learn all kinds of subjects, for example, statistics, management, database, etc., therefore, in contemporary society in the development of data mining technology is also more and more quickly, also more and more people like to use to integrate data mining and data warehouse technology, Once found that can use these data in the process of data mining, the value of the data warehouse technology will replace the data integration, data cleaning is to organize the error data or dirty data, therefore in the process of data mining must be combined with the data cleansing can let the data in the database to ensure the authenticity and validity. Therefore, in the development of data mining in China, there should be a lot of learning and improvement content, China should continue to establish and improve the data mining and data cleaning strategy research.

Keywords: Data mining; Data cleaning; Dirty data

引言：

在当今社会发展经济过程中，现代的企业不但要追求经济上面的发展，还要在决策方面积累经验，因此，在积累经验的时候，我们就应该采用到很多数据，数据的整合也成为现代社会发展的必然，因此，数据仓库成为整合数据的有效途径，当企业建立数据仓库之后，可以从企业的信息中筛选出自己想要的数据，数据仓库不单单是记录一个数据，而是记录多品种，多方面的数据，并且，在数据挖掘的过程中，推算出正确的决策方式，从而企业可以不断地积累决策的经验，但是在数据的挖掘过程中，我们也会碰到一些脏数据即没有用的数据，因此，我们应该使用数据清洗，将没有用的数据全都摒弃，而不是，接着留在数据库当中。

一、数据挖掘相关概念

数据挖掘指的是在自己建立的总的数据库当中，我们可以提取到自己需要的消息，主要用于将进行预测信息的技术数据挖掘，是一种专业挖掘信息以及预测的工具，他能够发现很多潜在信息，例如他可以，在一个公司的财务报表中，提取出其生产经营状况所能展现的利润，生产收入和生产成本等，这样就可以让企业的管理人员做出预测性的决策，当然，数据挖掘技术也不单单只有一种传统的技术，通常是采取假设的前提下，然后对所有的数据进行分析验证，这个假设是正确的，还是错误的，因此，数据挖掘现在只能够对所有的数据进行

作者简介：张祥飞（出生年—1994年），男，汉，河南省周口市，本科，职称/职务：软件工程师，现主要从事的工作或研究的方向：软件开发。

整理，并且做出预测性的分析，从而让公司做出正确的相应决策，因此，将其业务越做越好，帮助决策者更好地掌握市场的策略。

二、数据清洗的相关概念

数据清洗，从通俗的意义上来讲，也就是将数据的脏东西洗掉，指的就是在数据仓库中录入了很多数据，总会有一些是时间已经很久了的数据，或者说是已经没有用了的数据，这样我们可以将这些数据清洗掉，这些数据在数据库中，我们是很难发现，他们已经没有用了，因此，通常都是在数据挖掘过程中发现这些脏数据，并不是我们想要的数据，数据挖掘过程中，我们必须，演算出自己想要的东西，发现演算错误之后，我们就可以得到，在我们使用的所有数据里，有我们不需要使用的脏数据，因此，我们，要将这些脏数据给洗掉，这就是数据清洗的定义，但是呢，数据清洗的任务是把那些脏数据给清洗掉，清洗过程中我们需要，跟主管交代，确保自己不会将有用的数据给清洗掉，因此，在确认是否后再抽取那些没有用的数据及脏数据，清洗掉不符合要求的数据主要包括不完整的错误的重复的三大数据，数据清洗主要是在数据挖掘过程中发现数据的错误之后，然后从数据库当中将数据整理之后直接输出出去。

三、脏数据的类型及出现的原因

（一）脏数据的类型

1.缺失数据

造成数据缺失的原因主要有系统问题和人为问题等。假如出现了数据缺失情况，为了不影响数据分析产生结果的准确性，我们就应该将所缺数据及时补上，或者将空值排除在分析范围之外。

排除空值会减少数据分析的样本总量，这个时候可以选择性地纳入一些平均数、比例随机数等。若系统中还留有缺失数据的相关记录，可以通过系统再次引入，若系统中也没有这些数据记录，就只能通过补录或者直接放弃这部分数据来解决。

2.重复数据

相同的数据出现多次的情况相对而言更容易处理，因为只需要去除重复数据即可。但假如数据出现不完全重复的情况，例如某酒店VIP会员数据中，除了住址、姓名不一样，其余的大多数数据都是一样的，这种重复数据的处理就比较麻烦了。假如数据中有时间、日期，仍然可以以此作为判断标准来解决，但假如没有时间、日期这些数据，就只能通过人工筛选来处理。

3.错误数据

错误数据的产生经常是由于在录入数据之前，没有按照规定流程走程序。例如异常值，某个产品价格为1到100元，而统计中偏偏出现200这个值；例如格式错误，将天气录成了文字格式；例如数据的不统一，关于天津的记录有天津、tianjin。对于异常值，可以通过限定区间的方法进行排除；对于格式错误，需要通过系统内部逻辑结构进行查找；对于数据不统一，无法从系统方面去解决，因为它并不属于真正的“错误”，系统并不能判断出天津和tianjin属于同一“事物”，因此只能通过人工干预的方法，做出匹配规则，用规则表去关联原始表。例如，一旦出现tianjin这个数据就直接匹配到天津[1]。

4.不可用数据

有些数据虽然正确但却无法使用。例如地址为“上海浦东新区”，想要对“区”级别的数据进行分析时，还需要将“浦东”拆出来。这种情况的解决方案只能用关键词匹配的方法，而且不一定能够得到完美解决。

（二）脏数据出现的原因

“脏”数据又是什么呢？通俗来说，它是因数据重复录入、共同处理等不规范操作而产生的混乱、无效数据。这些数据不能为企业带来价值，反而会占据存储空间，浪费企业的资源。因此，这些数据被称为“脏”数据，不仅没有价值，还会“污染”其他的数据。某些“脏”数据还可能给企业带来重大损失。曾经有一家保险公司，把客户的资料存储在数据库中，并进行了如下规定：在存入新的数据之前，要对数据库进行检索，以查看其中是否存在相关记录。然而，一些数据员偷懒，擅自跳过搜索环节，直接存入了新的数据，导致数据的重复录入。久而久之，系统运行越来越缓慢，搜索结果越来越不准确，最终数据库完全失灵，给公司造成巨大的经济损失。这个时候，保险公司才如梦初醒，决定解决这个问题。公司花费了一个星期的时间，将这些积存在数据库中的“脏”数据全部清除。当数据出现问题的时候，苦心构建的数据库就失去了原有价值。正因如此，处理“脏”数据的工作就变得十分重要，而且越早开始越好。因此，我们有必要了解一下“脏”数据的种类[2]。

四、数据清洗的定义与对象

（一）数据清洗的定义

数据清洗到现在都没有一个固定的定义，在数据库当中，数据清洗是由于数据挖掘时产生的数据影响而产生的词，在数据挖掘过程中，会出现一些数据的相关错误，因此，这些数据就应该采用数据清洗的功能，将其摒弃，因此，数据清洗不同的地方，也有不同的定义。简单来说，数据清洗就是为了让下次不犯这一次同样的

错误，因此，在数据挖掘过程中发现的数据问题，就应该把这个脏的数据给摒弃掉，因此，数据清洗的主要作用就是为了让数据更加准确，更加明了。

（二）数据清洗的对象

建立数据库的主要目的就是为了在数据发掘过程中利用数据发掘工具进行分析，得到数据结果，为了保证数据分析得到的结果是一个非常准确的数据，就保证所采用的数据都是干净的数据，所以数据清洗就显得尤为重要，数据让数据变得更加准确，可以统一数据的计算方法和格式，这样就能减少在数据分析过程中会产生的各种问题，从而提高他的效率，清洗数据的对象主要有三方面，一是缺乏值，二是重复值，三是异常值等。

缺乏值通常指的是，在数据挖掘过程中，数据库中总会有一些没有录到的数据，因此，少了这一个数据值，我们可能无法特别准确的判断数据的分析，因此，这一个缺乏的值就叫缺乏值，包括说分组在数据挖掘过程中缺少了某些分组，这个分组也统一叫缺乏值，因此，这样看来，缺乏了值就会对数据分析有一定的影响，对于某一个样本，某一值缺乏太多，我们应该将其全部删掉，这样我们可以将其数据分析的不准确性降到最低，对于这些样本，我们也应该去积极采集这些缺乏值得由于数据库中没有录到我们应该及时采集。

其次就是异常值，这里的异常值指的就是在数据录入过程中，数据录的不够准确，没有通过取证的方式将其数据准确性确定，因此，在数据不准确的情况下，我们所判断分析得到的结果就是不准的，通常我们用平均值的方式来判断这个值是否是异常值，如若算出来的平均值偏差超过了测定偏差的两倍，我们就判断这个值是不准确的，在对于异常值，一般我们不会做处理，当然，如果说对异常值做了处理，这样子可以省去功夫。

最后是重复值，重复值通常意义上来说，就是在数据库中，相同的数据重复值一般会分为两种，一种是完全相同的数据，出现了两个数据库的结果，另一种数据则是在不同的数据库当中出现了相同的数据，这可能是在处理过程中没有做好数据录入的准备，去除这些重复值，可以采用驱虫和去除的方法。

关于数据清洗的对象和内容就是以上，这些通常来讲，数据清理的主要工作就是除去数据中存在的异常值，缺乏值和重复值，这些没有用的数据可能会对大家在数据仓库中数据分析时产生一定影响，数据分析可能会得到准确的结果，因此，大家在处理数据前一定要确保录入的数据是准确的，并且删除数据时应该看清楚所除数据是否是保存好自己的初始数据[3]。

五、数据清洗结的方法

数据清洗有很多种方法，一般来说，大家都采用属性及异常数据的清洗，和记录及异常数据的清洗来清洗脏数据。数据清理主要指的就是在数据库中将重复的记录去除，并且将其它的数据转换成可用数据，不全去除而是只去除一个数据，数据清理可以用模型进行清理，也可以通过去除或去除的方式进行清理，但是都要使用一定的方式才能将重复的数据清理掉。数据库中有大多数数据都是可以用的，但是少部分数据需要进行数据清理，数据清理从多方面来展示怎么处理数据的问题？数据处理一般针对具体的数据很难归纳有统一的流程处理，因此对不同的数据可以有不一样的数据清理方法。

（一）解决不完整数据的方法

不完整数据，其实缺乏的数据对于缺乏的数据呢，通常要进行手工的输入或者是进行手工的清理，当然，有一些缺乏值是通过一些数据源得来的，即是要通过一些公式演算得来，这样我们可以通过用演算公式得到的值来代替缺乏的值，这样我们也就达到了清理的目的。举例，如果有一个值是计算其平均值与当季最大值的比较，我们就可以不需要用原有数据，而是只需要用其平均值，即可得出结果[4]。

（二）错误值的检测及解决方法

对于错误的值，我们应该如何摒弃呢？我们应该在使用回归方程等计算公式之后，我们可以发现，其数据到底是不是异常值？如果说是常值的话，我们可以使用简单的规则库来检查所有的数据字，或者使用外部数据检测来达到清理数据的目的。

（三）重复记录的检测及消除方法

对于重复数据，我们可以采用，去重的方法来进行消除，或是将两条相同的数据进行合并，都可以达到重复记录消除的目的。

（四）不一致性的检测及解决方法

如果有很多数据都是不一致性的，我们可以通过采用其他数据源的数据库来进行比较，通过比较分析数据，我们可以发现，其数据是否具有完整性，从而使得数据最后的结果是保持一致的。

六、数据清洗研究的不足与展望

在目前的发展情况下，我国的数据清洗技术还是非常的不健全的，相较于国外对数据清洗的研究十分不成熟，中文数据的分析比较少，国内在对于数据清洗方面主要集中在算法方面，原创性的东西还比较少，因此，取得的效果也不多，在对数据清洗的分析过程中，我们还存在着，许多的发展空间有很大的前景和商讨价值。

其中对于数据清洗不足之处主要表现在以下几个方面：

（1）在数据清理的研究上，我们主要是会采取西方的数据，我国自己的数据没有得到广泛的采用，因此中文的数据清理方法还没有得到有效的重视。

（2）在现有的研究基础上，数据清洗主要还是针对在表层的数据上，比如说在数值上有着很强大的要求，但是在模式上的数据清洗研究的非常少，因此，在不同层面上的数据清洗发展的特别规律的[5]。

（3）在数据库中，还存在着许多重复数据的问题，在数据清洗过程中，我们对于重复数据的识别率是非常低的，在记录数据时耗时非常多，记录数据非常枯燥繁琐，因此，其对重复数据的识别工程特别大。

（4）在数据清理的过程之中，我们没有一个结构化的处理，单单的采取原来的旧方式来处理重复的数据，而清理的对象也是用之前的架构和方法来清理，并没有创新性。

（5）数据清洗工具有很多种，但是我国只采用了描述型的数据来清理，这样不仅不能很有效的取得数据清理的结果，还可能会导致有部分的数据没有得到清理，因此，我国应该采取不同的数据清理工具，来进行数据的结构化清理。

（6）现在的数据清理方式主要是面向比较特定的领域，应该采取多方领域，多模式发展。

由于我国数据清洗存在许多不足之处，因此数据清洗未来主要的研究方向有：

（1）我们在开发外国数据清理时，应该对中文的数据进行，合理开发在发展中文数据清理工具时，我们应该对其积极研究，将他的作用发展到好的阶段[6]。

（2）在数据挖掘的方式和方法上，我们应该在数据清理的多方领域做应用的研究，将其深入发展。

（3）我们在数据清理的过程中，重复值的识别率是非常低的，因此，在寻找重复值时，我们的工作量是非常大的，耗时时间特别长，因此，我们应该提高重复记录识别的效率。

（4）在结构化数据清洗过程中，我们也会做到有一些数据根本就没有得到数据清洗，所以我们应该采取非结构化数据的清洗，让每一个数据都可以得到清洗，将数据库中所有的数据都整理到位。

（5）在数据清洗过程中，我们所使用到的工具是有互操作性的，我们应该好好的利用其互操作性的功能，清理不一样的数据可以采用不一样的工具，积极使用每一个工具，将数据清理的效率达到最高。

七、结语

因此，我国在发展数据挖掘过程中，还应该有很多学习和改善的内容，我国应该不断建立健全数据挖掘和数据清洗的各项策略研究。在数据清洗时，我们应该认识自身的不足之处，并且将其加以改正，对未来的研究方向也提出了相应的要求，相信我国未来数据清洗研究一定可以健康有效的充分发展。

参考文献：

[1]赵巧稚.基于模糊神经网络的城市污水处理过程数据清洗方法的研究及应用[D].北京：北京工业大学，2020.

[2]邹同华，高云鹏，伊慧娟，等.基于Thompson tau-四分位和多点插值的风电功率异常数据处理[J].电力系统自动化，2020，44（15）：156-162.DOI：10.7500/AEPS20191231003.

[3]刘振鹏，苏楠，秦益文，等.FS-CRF：基于特征切分与级联随机森林的异常点检测模型[J].计算机科学，2020，47（8）：185-188.DOI：10.11896/jsjkx.190600162.

[4]杨光，吴明芬，李敬民.数据迁移与清洗的策略研究及其在政务基础数据的应用[J].五邑大学学报（自然科学版），2021，35（1）：55-61.DOI：10.3969/j.issn.1006-7302.2021.01.011.

[5罗琨.ETL技术在提高统一社会信用代码数据质量中的应用研究[J].标准科学，2020（6）：103-108.DOI：10.3969/j.issn.1674-5698.2020.06.018.

[6] 陈新月.基于并行计算的水质时间序列数据清洗平台的研究与实现[D].北京：北京工业大学，2020.