(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 107741990 B
(45)授权公告日 2023.05.16

(21)申请号 201711059055.X

(22)申请日 2017.11.01

(65)同一申请的已公布的文献号
申请公布号 CN 107741990 A

(43)申请公布日 2018.02.27

(73)专利权人 深圳汇生通科技股份有限公司
地址 518000 广东省深圳市南山区南山街道前海路3101-90号振业国际商务中心3008

(72)发明人 高霞光 刘军

(74)专利代理机构 深圳市科冠知识产权代理有限公司 44355
专利代理师 王海骏

(51)Int.Cl.
G06F 16/215(2019.01)

(56)对比文件
CN 103279863 A,2013.09.04
CN 107025293 A,2017.08.08
CN 101329731 A,2008.12.24
CN 106294480 A,2017.01.04
CN 104636741 A,2015.05.20
CN 106776703 A,2017.05.31
CN 106021196 A,2016.10.12
US 5862400 A,1999.01.19

审查员 王一

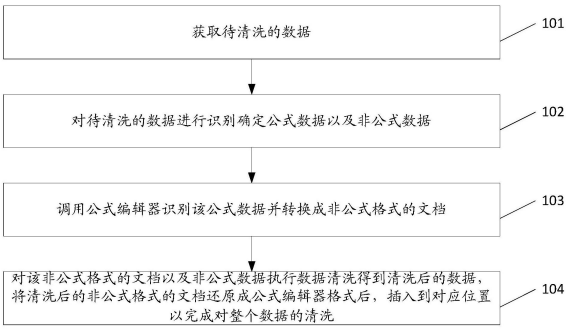
权利要求书1页 说明书5页 附图1页

(54)发明名称

数据清洗整合方法及系统

(57)摘要

本发明公开了一种数据清洗整合方法及系统，所述方法包括如下步骤：获取待清洗的数据；对待清洗的数据进行识别确定公式数据以及非公式数据；调用公式编辑器识别该公式数据并转换成非公式格式的文档；对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据，将清洗后的非公式格式的文档还原成公式编辑器格式后，插入到对应位置以完成对整个数据的清洗。本发明提供的技术方案能够对公式进行处理的优点。



CN 107741990 B

1.一种数据清洗整合方法,其特征在于,所述方法包括如下步骤:

获取待清洗的数据;对待清洗的数据进行识别确定公式数据以及非公式数据;

调用公式编辑器识别该公式数据并转换成非公式格式的文档;

对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据,将清洗后的非公式格式的文档还原成公式编辑器格式后,插入到对应位置以完成对整个数据的清洗。

2.根据权利要求1所述的方法,其特征在于,所述调用公式编辑器识别该公式数据并转换成非公式格式的文档,具体包括:

提取该公式数据中除了符号以外的数据以及符号的顺序,将符号以外的数据转换成非公式数据。

3.根据权利要求1所述的方法,其特征在于,所述对待清洗的数据进行识别确定公式数据以及非公式数据,包括:

对该数据进行识别确定数据的格式,如该格式为非文档格式,确定为公式数据。

4.根据权利要求1所述的方法,其特征在于,所述方法还包括:

所述数据清洗包括:无效值和缺失值的处理或一致性检查。

5.一种数据清洗整合系统,其特征在于,所述系统包括:

获取单元,用于获取待清洗的数据;

处理单元,用于对待清洗的数据进行识别确定公式数据以及非公式数据;调用公式编辑器识别该公式数据并转换成非公式格式的文档;对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据,将清洗后的非公式格式的文档还原成公式编辑器格式后,插入到对应位置以完成对整个数据的清洗。

6.根据权利要求5所述的系统,其特征在于,所述处理单元,还用提取该公式数据中除了符号以外的数据以及符号的顺序,将符号以外的数据转换成非公式数据。

7.根据权利要求5所述的系统,其特征在于,所述处理单元,具体用于

对该数据进行识别确定数据的格式,如该格式为非文档格式,确定为公式数据。

8.根据权利要求5所述的系统,其特征在于,

所述数据清洗包括:无效值和缺失值的处理或一致性检查。

9.一种计算机可读存储介质,其特征在于,其存储用于电子数据交换的计算机程序,其中,所述计算机程序使得计算机执行如权利要求1-4任一项所述的方法。

数据清洗整合方法及系统

技术领域

[0001] 本发明涉及数据处理领域，尤其涉及一种数据清洗整合方法及系统。

背景技术

[0002] 数据清洗(Data cleaning)-对数据进行重新审查和校验的过程，目的在于删除重复信息、纠正存在的错误，并提供数据一致性。

[0003] 数据清洗从名字上也看的出就是把“脏”的“洗掉”，指发现并纠正数据文件中可识别的错误的最后一道程序，包括检查数据一致性，处理无效值和缺失值等。因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。我们要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。而数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要是有不完整的数据、错误的数据、重复的数据三大类。数据清洗是与问卷审核不同。现有的数据清洗无法对公式进行整合调整。

发明内容

[0004] 本申请提供一种数据清洗整合方法。其解决现有技术的技术方案无法清洗公式的缺点。

[0005] 一方面，提供一种数据清洗整合方法，所述方法包括如下步骤：

[0006] 获取待清洗的数据；对待清洗的数据进行识别确定公式数据以及非公式数据；

[0007] 调用公式编辑器识别该公式数据并转换成非公式格式的文档；

[0008] 对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据，将清洗后的非公式格式的文档还原成公式编辑器格式后，插入到对应位置以完成对整个数据的清洗。

[0009] 可选的，所述调用公式编辑器识别该公式数据并转换成非公式格式的文档，具体包括：

[0010] 提取该公式数据中除了符号以外的数据以及符号的顺序，将符号以外的数据转换成非公式数据。

[0011] 可选的，所述对待清洗的数据进行识别确定公式数据以及非公式数据，包括：

[0012] 对该数据进行识别确定数据的格式，如该格式为非文档格式，确定为公式数据。

[0013] 可选的，所述方法还包括：

[0014] 所述数据清理包括：无效值和缺失值的处理或一致性检查。

[0015] 第二方面，提供一种数据清洗整合系统，所述系统包括：

[0016] 获取单元，用于获取待清洗的数据；

[0017] 处理单元，用于对待清洗的数据进行识别确定公式数据以及非公式数据；调用公

式编辑器识别该公式数据并转换成非公式格式的文档;对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据,将清洗后的非公式格式的文档还原成公式编辑器格式后,插入到对应位置以完成对整个数据的清洗。

[0018] 可选的,所述处理单元,还用提取该公式数据中除了符号以外的数据以及符号的顺序,将符号以外的数据转换成非公式数据。

[0019] 可选的,所述处理单元,具体用于

[0020] 对该数据进行识别确定数据的格式,如该格式为非文档格式,确定为公式数据。

[0021] 可选的,所述数据清理包括:无效值和缺失值的处理或一致性检查。

[0022] 第三方面,提供一种计算机程序产品,所述计算机程序产品包括存储了计算机程序的非瞬时性计算机可读存储介质,所述计算机程序可操作来使计算机执行第一方面所述的方法。

[0023] 一种计算机可读存储介质,其存储用于电子数据交换的计算机程序,其中,所述计算机程序使得计算机执行第一方面所述的方法。

[0024] 本发明提供的技术方案对公式进行转换成非公式数据,然后清洗以后在转换成公式数据,从而实现对公式数据执行清洗。

附图说明

[0025] 为了更清楚地说明本发明实施例的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图是本发明的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0026] 图1为本发明第一较佳实施方式提供的一种数据清洗整合方法的流程图;

[0027] 图2为本发明第二较佳实施方式提供的一种数据清洗整合系统的结构图。

具体实施方式

[0028] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0029] 请参考图1,图1是本发明第一较佳实施方式提出的一种数据清洗整合方法,该方法如图1所示,包括如下步骤:

[0030] 步骤S101、获取待清洗的数据;

[0031] 步骤S102、对待清洗的数据进行识别确定公式数据以及非公式数据;

[0032] 上述步骤的实现方法具体可以为,对该数据进行识别确定数据的格式,如该格式为非文档格式,确定为公式数据。。

[0033] 步骤S103、调用公式编辑器识别该公式数据并转换成非公式格式的文档。

[0034] 上述步骤的实现方法可以为,提取该公式数据中除了符号以外的数据以及符号的顺序,将符号以外的数据转换成非公式数据。

[0035] 步骤S104、对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据,将清洗后的非公式格式的文档还原成公式编辑器格式后,插入到对应位置以完成对整

个数据的清洗。

[0036] 上述清洗的数据可以采用的方法具体可以为：

[0037] 一致性检查

[0038] 一致性检查(consistency check)是根据每个变量的合理取值范围和相互关系，检查数据是否合乎要求，发现超出正常范围、逻辑上不合理或者相互矛盾的数据。例如，用1-7级量表测量的变量出现了0值，体重出现了负数，都应视为超出正常值域范围。SPSS、SAS、和Excel等计算机软件都能够根据定义的取值范围，自动识别每个超出范围的变量值。具有逻辑上不一致性的答案可能以多种形式出现：例如，许多调查对象说自己开车上班，又报告没有汽车；或者调查对象报告自己是某品牌的重度购买者和使用者，但同时又在熟悉程度量表上给了很低的分值。发现不一致时，要列出问卷序号、记录序号、变量名称、错误类别等，便于进一步核对和纠正。

[0039] 无效值和缺失值的处理

[0040] 由于调查、编码和录入误差，数据中可能存在一些无效值和缺失值，需要给予适当的处理。常用的处理方法有：估算，整例删除，变量删除和成对删除。

[0041] 估算(estimation)。最简单的办法就是用某个变量的样本均值、中位数或众数代替无效值和缺失值。这种办法简单，但没有充分考虑数据中已有的信息，误差可能较大。另一种办法就是根据调查对象对其他问题的答案，通过变量之间的相关分析或逻辑推论进行估计。例如，某一产品的拥有情况可能与家庭收入有关，可以根据调查对象的家庭收入推算拥有这一产品的可能性。

[0042] 整例删除(case wise deletion)是剔除含有缺失值的样本。由于很多问卷都可能存在缺失值，这种做法的结果可能导致有效样本量大大减少，无法充分利用已经收集到的数据。因此，只适合关键变量缺失，或者含有无效值或缺失值的样本比重很小的情况。

[0043] 变量删除(variable deletion)。如果某一变量的无效值和缺失值很多，而且该变量对于所研究的问题不是特别重要，则可以考虑将该变量删除。这种做法减少了供分析用的变量数目，但没有改变样本量。

[0044] 成对删除(pair wise deletion)是用一个特殊码(通常是9、99、999等)代表无效值和缺失值，同时保留数据集中的全部变量和样本。但是，在具体计算时只采用有完整答案的样本，因而不同的分析因涉及的变量不同，其有效样本量也会有所不同。这是一种保守的处理方法，最大限度地保留了数据集中的可用信息。

[0045] 采用不同的处理方法可能对分析结果产生影响，尤其是当缺失值的出现并非随机且变量之间明显相关时。因此，在调查中应当尽量避免出现无效值和缺失值，保证数据的完整性。

[0046] 残缺数据

[0047] 这一类数据主要是一些应该有的信息缺失，如供应商的名称、分公司的名称、客户的区域信息缺失、业务系统中主表与明细表不能匹配等。对于这一类数据过滤出来，按缺失的内容分别写入不同Excel文件向客户提交，要求在规定的时间内补全。补全后才写入数据仓库。

[0048] 错误数据

[0049] 这一类错误产生的原因是业务系统不够健全，在接收输入后没有进行判断直接写

入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车操作、日期格式不正确、日期越界等。这一类数据也要分类，对于类似于全角字符、数据前后有不可见字符的问题，只能通过写SQL语句的方式找出来，然后要求客户在业务系统修正之后抽取。日期格式不正确的或者是日期越界的这一类错误会导致ETL运行失败，这一类错误需要去业务系统数据库用SQL的方式挑出来，交给业务主管部门要求限期修正，修正之后再抽取。

[0050] 重复数据

[0051] 对于这一类数据——特别是维表中会出现这种情况——将重复数据记录的所有字段导出来，让客户确认并整理。

[0052] 数据清洗是一个反复的过程，不可能在几天内完成，只有不断的发现问题，解决问题。对于是否过滤，是否修正一般要求客户确认，对于过滤掉的数据，写入Excel文件或者将过滤数据写入数据表，在ETL开发的初期可以每天向业务单位发送过滤数据的邮件，促使他们尽快地修正错误，同时也可以做为将来验证数据的依据。数据清洗需要注意的是不要将有用的数据过滤掉，对于每个过滤规则认真进行验证，并要用户确认。

[0053] 解决不完整数据(即值缺失)的方法

[0054] 大多数情况下，缺失的值必须手工填入(即手工清理)。当然，某些缺失值可以从本数据源或其它数据源推导出来，这就可以用平均值、最大值、最小值或更为复杂的概率估计代替缺失的值，从而达到清理的目的。

[0055] 错误值的检测及解决方法

[0056] 用统计分析的方法识别可能的错误值或异常值，如偏差分析、识别不遵守分布或回归方程的值，也可以用简单规则库(常识性规则、业务特定规则等)检查数据值，或使用不同属性间的约束、外部的数据来检测和清理数据。

[0057] 重复记录的检测及消除方法

[0058] 数据库中属性值相同的记录被认为是重复记录，通过判断记录间的属性值是否相等来检测记录是否相等，相等的记录合并为一条记录(即合并/清除)。合并/清除是消重的基本方法。

[0059] 不一致性(数据源内部及数据源之间)的检测及解决方法

[0060] 从多数据源集成的数据可能有语义冲突，可定义完整性约束用于检测不一致性，也可通过分析数据发现联系，从而使得数据保持一致。目前开发的数据清理工具大致可分为三类。

[0061] 数据清洗工具使用领域特有的知识(如，邮政地址)对数据作清洗。它们通常采用语法分析和模糊匹配技术完成对多数据源数据的清理。某些工具可以指明源的“相对清洁程度”。工具Integrity和Trillum属于这一类。

[0062] 本发明提供的技术方案对公式进行转换成非公式数据，然后清洗以后在转换成公式数据，从而实现对公式数据执行清洗。

[0063] 请参考图2，图2是本发明第二较佳实施方式提出的一种数据清洗整合系统，所述系统包括：

[0064] 获取单元201，用于获取待清洗的数据；

[0065] 处理单元202，用于对待清洗的数据进行识别确定公式数据以及非公式数据；调用

公式编辑器识别该公式数据并转换成非公式格式的文档;对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据,将清洗后的非公式格式的文档还原成公式编辑器格式后,插入到对应位置以完成对整个数据的清洗。

[0066] 可选的,所述处理单元,还用提取该公式数据中除了符号以外的数据以及符号的顺序,将符号以外的数据转换成非公式数据。

[0067] 可选的,所述处理单元,具体用于

[0068] 对该数据进行识别确定数据的格式,如该格式为非文档格式,确定为公式数据。

[0069] 可选的,所述数据清理包括:无效值和缺失值的处理或一致性检查。

[0070] 需要说明的是,对于前述的各个方法实施例,为了简单描述,故将其都表述为一系列的动作组合,但是本领域技术人员应该知悉,本发明并不受所描述的动作顺序的限制,因为依据本发明,某一些步骤可以采用其他顺序或者同时进行。其次,本领域技术人员也应该知悉,说明书中所描述的实施例均属于优选实施例,所涉及的动作和模块并不一定是本发明所必须的。

[0071] 在上述实施例中,对各个实施例的描述都各有侧重,某个实施例中没有详细描述的部分,可以参见其他实施例的相关描述。

[0072] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成,该程序可以存储于一计算机可读存储介质中,存储介质可以包括:闪存盘、只读存储器(英文:Read-Only Memory,简称:ROM)、随机存取器(英文:Random Access Memory,简称:RAM)、磁盘或光盘等。

[0073] 以上对本发明实施例所提供的内容下载方法及相关设备、系统进行了详细介绍,本文中应用了具体个例对本发明的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本发明的限制。

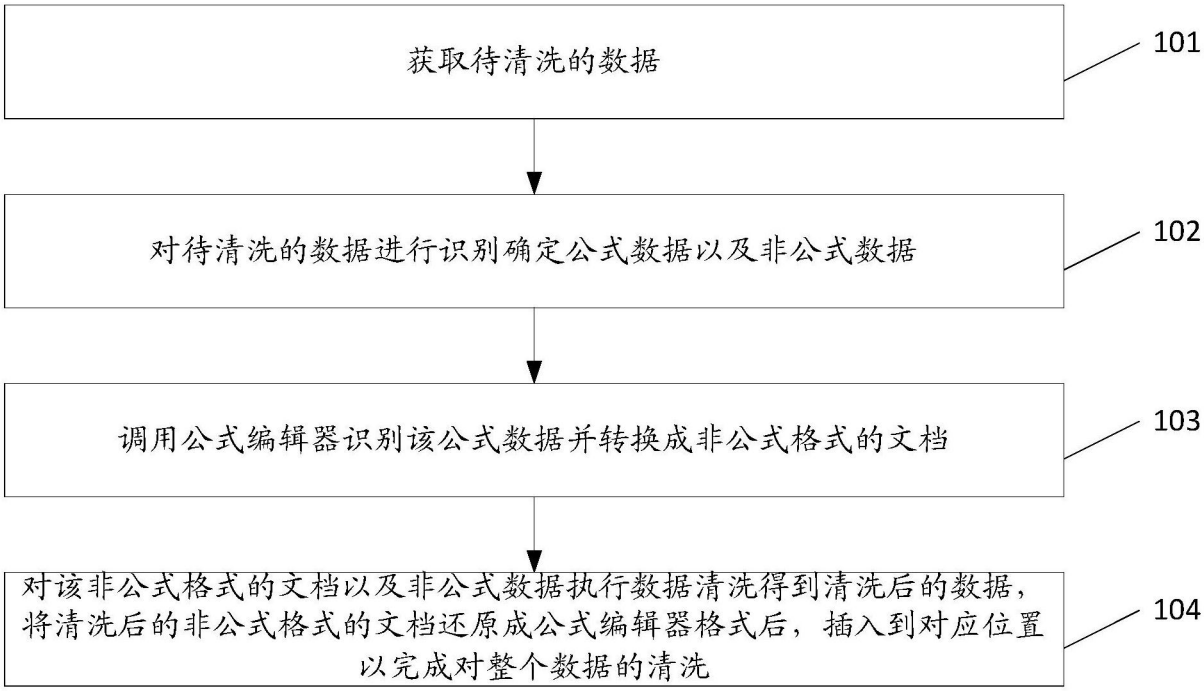

获取待清洗的数据
101
对待清洗的数据进行识别确定公式数据以及非公式数据
102
调用公式编辑器识别该公式数据并转换成非公式格式的文档
103
对该非公式格式的文档以及非公式数据执行数据清洗得到清洗后的数据，将清洗后的非公式格式的文档还原成公式编辑器格式后，插入到对应位置以完成对整个数据的清洗
104

图1

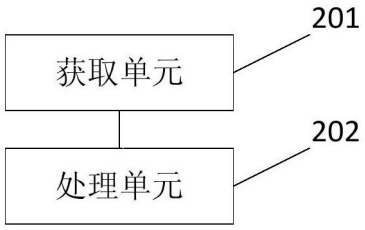

201
获取单元
202
处理单元

图2