利用实时性来修复不一致和不完整的数据

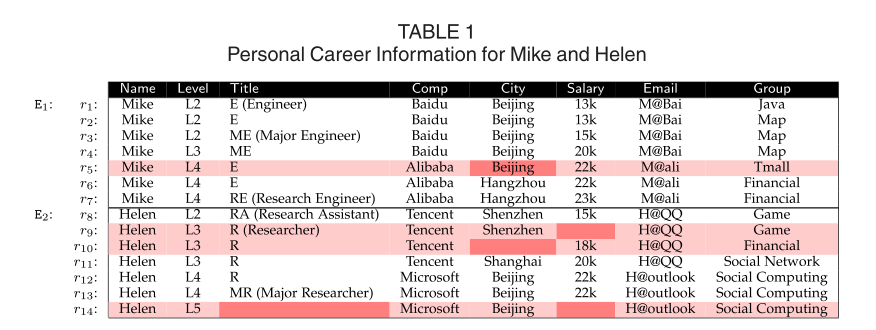
摘要--今天，数据质量在大数据管理中起着关键作用。随着各种来源的数据的爆炸性增长，数据的质量面临着多种问题。受此启发，我们在本文中研究了关于不完全性和不一致性的多种数据清洗的货币推理和判断。我们引入了一个名为Imp3C的四步框架，用于在没有时间戳的不完整和不一致的数据中检测错误和提高质量。我们实现了一个综合的货币确定方法，根据货币约束计算图元之间的货币顺序。因此，考虑到时间上的影响，不一致的数据和缺失的数值被有效地修复了。出于有效性和效率的考虑，我们在不完全性修复之前进行了不一致修复。我们定义了一个与货币相关的一致性距离指标，以更准确地衡量脏图元和干净图元之间的相似性。此外，在缺失归因训练过程中，货币顺序被视为一个重要特征。通过案例研究，详细介绍了解决方案的算法。在三个现实生活中的数据集上进行的全面实验验证了我们的方法Imp3C改善了有多种质量问题的数据修复的性能。Imp3C优于现有的先进方法，特别是在具有复杂货币订单的数据集中。

1 简介

DATA质量在以数据为中心的应用中起着关键作用[1]。数据质量问题相当严重，困扰着数据交易步骤（如采集、复制、查询）。具体来说，货币、一致性和完整性是数据清理的三个重要问题[2]。各种信息系统以不同的格式或语义存储数据。这可能会导致多源数据整合中昂贵的一致性问题。此外，由于信息系统的完整性标准不完善，数据库中的图元可能存在缺失值[3]。更糟糕的是，数据更新的频率很低，在数据源之间松散或不精确的数据复制功能下，时间戳的缺失会使数据在某种程度上过期[4]。这三个问题导致了数据的低可靠性，从而增加了数据应用中的混乱。因此，低质量的数据可能会对许多领域造成负面影响。研究人员在数据修复方面已经取得了很大的进展，特别是在一致性和完整性方面。人们承认，在修复过程中，一致性和完整性问题可能会相互影响，而不是完全孤立的[2]，[5]，有时这两种方法是互补的。我们发现，货币问题也严重影响了不一致和不完整值的修复。有了数据库中精确的时间戳，诸如实体解析和基于约束的数据修复等数据清理任务就可以通过最先进的数据清理方法[6]和系统[7]来解决了。然而，时间性数据中不完整的时间戳使得这个问题在检测和修复上都面临挑战。由于在客户关系管理（CRM）[8]和人力资源管理（HRM）[5],[9]等信息系统中使用最新的数据和了解记录的及时性是非常重要的，在不一致和不完整的低质量数据集中，不完整的时间戳使得这个问题在检测和修复上都面临挑战。我们提出了一个行业人才数据库的激励例子，以说明现有的数据修复方法在不利用货币问题的情况下的缺点。

1. 表1显示了一个行业人才数据库中记录的几个图元，描述了两个人（实体），Mike和Helen，有8个属性。级别是行业认可的职业等级，而职称是员工的职位。城市记录了公司所处的城市。缩略语用于当前的专业电子邮件中。由于数据来自多个来源，时间戳缺失。属性值都存在不一致和不完整的问题。

表1用红色标记了4个图元中的错误值。对于Mike来说，r5描述Mike在Comp2（杭州）工作。同时，该城市是北京。这导致了混淆，我们可以得出结论，r5中的城市，甚至是公司和集团都存在不一致的值。对于Helen来说，r9、r10和r14包含缺失的值。我们很难推断她是什么时候开始在上海工作的，以及她现在的工资是多少。表2显示了现有数据修复方法的可选修复方案。对于Mike，我们可以给r9的缺失工资一个相对干净的值15k，参考其最类似的记录r8，但在修复其他错误时，事情并不简单。Mike工作的公司和集团都与r5中的城市不一致。根据[10]中提出的最小修复原则，用r4的属性值修复r5是可能的。然而，在r5的时候，Mike实际上已经开始在Comp2工作，这意味着r5比r4更有时效性。因此，在不考虑时间问题的情况下，这种修复是很差的。对于Helen来说，根据[11]的编辑距离distðr9; r10Þ ¼ distðr10; r11Þ，很难区分哪个更接近r10，而且它在r9; r10和r11之间也没有呈现出货币差异。同样，由于r12; r14Þ和r13; r14Þ的距离相等，用R或MR来修复r14似乎也没有区别。从上述情况来看，如果没有可用的时间戳的指导，很难清理不一致的和不完整的值。如果简单地用它们最相似的图元的值来修复，很可能得到错误的修复结果，甚至在数据应用中遇到严重的错误。因此，在货币、一致性和完整性（3C）方面共同修复数据质量问题是很有必要的。然而，混合质量清洗问题面临着挑战。首先，随着属性值随时间的变化和发展，图元中的时间特征会影响修复的准确性。此外，由于一些整体的基本问题被称为计算上的困难[4]，[10]，多错误的数据修复使其面临更大的挑战。更糟糕的是，修复一个错误可能会导致任何其他类型的错误。如果没有一个复杂的方法，由于数据清洗造成的反复错误修复步骤，修复脏数据的成本很高。到目前为止，关于用数据货币清理多种错误的工作仍然是不充分的。当时间戳不可用时，货币顺序很难被确定。现有的货币修复方法大多依赖于明确的时间戳（例如[9]），很少有作品为缺乏有效时间戳的数据提供可行的方法。受此启发，我们用数据货币订单来提高数据修复的一致性和完整性的质量。我们继续说明例1中的货币决定在数据修复中的好处。表2用绿色显示了表1中图元的更好的修复。我们推断出一个货币顺序，即在现实世界中，一个公司的员工的头衔是不断增加的。因此，当Mike在同一公司工作时，他的头衔只能从E变成ME。同样地，工资总是单调地增加。因此，预计r5比r4更具时效性。r5½City? 应该被修复。错误的发生可能是因为数据库更新和现实世界的变化之间存在延迟。此外，如果电子邮件不能得到很好的修复，员工和公司都将遭受损失。对于Helen，可以推断出r10½City? ¼上海，根据CFD: ð½Comp? ¼ Comp3 ^ ½Group? ¼ FinancialÞ ! ½城市？¼ 上海Þ。它揭示了Helen已经将她的工作转到了L3的上海的金融组。这种修复提高了她职业信息的准确性。通过一个货币顺序。R?Title MR，我们知道Helen在L4成为MR，并且r13比r12更有时效性。根据另一个货币顺序。L4 ?Level L5，r14是目前最新鲜的元组。它缺失的头衔和工资应该用最新的值来填补，即分别为L5和22k。这表明Helen在L5的工资不低于22k，是她组里的一个MR。这些案例说明了脏数据中的复杂情况，以及交互法在3C数据清洗中的必要性。从表2中可以看出，这三个问题的结合对数据清洗的准确性提高做出了贡献。



贡献。我们将我们的贡献总结如下。1）我们提出了一种全面的一致性、完整性和货币性的数据修复方法，命名为Imp3C。据我们所知，这是第一个在没有可靠时间戳的情况下解决不一致修复和缺失归因的数据清理的研究。2）我们提出了一种货币图谱构建和货币顺序计算的综合方法，具有货币约束。在时间戳无效的情况下，它实现了可靠的时间相关替换，以便充分使用数据库中的货币信息。此外，我们根据货币图谱定量地计算货币订单对的可靠性指标。通过这种方式，可以发现和唤醒图元中的内部时间知识，以最大限度地提高修复的有效性。3）我们定义了脏数据和标准数据之间的货币一致性距离度量，以便与CFD和货币订单一起修复不一致的属性。此外，我们提出了基于Naıve Bayes的缺失值估算解决方案，其中货币订单被认为是训练过程中的一个关键特征。4）我们在三个现实生活中的数据集上进行了全面的实验。实验结果证明，Imp3C的修复策略是有效的。Imp3C在修复多错误方面优于现有方法，特别是在货币频繁演变和变化的数据中。组织结构。本文的其余部分组织如下。第2节讨论了基本定义和我们方法的概述。第3节介绍了货币图的构建方法，第4节讨论了货币顺序确定算法。第5节（即第6节）介绍了不一致修复（即缺失值归纳）方法。第7节报告了实验研究。第8节回顾了相关工作，第9节给出了结论。

2 概述

在本节中，我们首先在第2.1节中介绍背景和基本定义，然后在第2.2节中提出方法框架。

2.1 前言

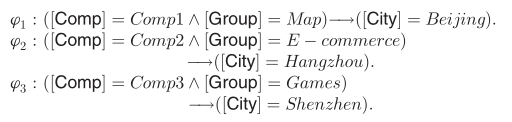
在下文中，我们将D表示为一个数据集，其中的图元满足关系模式R ¼ frID; eID; A1; . . ; Ang。rID是图元的ID号，eID是实体的ID号。attrðRÞ是R的属性集，ri½Aj? 是ri在Aj上的属性值，Aj2attrðRÞÞ。我们假设对一个属性Aj的值有一个部分顺序，我们用"？"来表示。D中每个属性的货币顺序Ak是R中图元的一个部分顺序，其特性是它只与具有相同eID的图元相关，即ri ?Ak rj暗示ri½eID? ¼ rj½eID?。接下来，我们引入货币约束（简称CC），它将属性Ak的值和它们的顺序与定义1中的货币顺序?Ak联系起来。本文采用的货币约束的语义参考了[10]。

定义1. R的货币约束c被表示为8s; t : R?ðs½eID? ¼ t½eID? ^ fÞ?!s ?Ak t?。设ri; rj为元组变量，表示R的一个元组，f是一个来自：（1）ri½Ah? ¼ c ðresp: ri½Ah? 6¼ cÞ，其中c是一个常数；(2) ri½Ah? op rj½Ah?; o p 2 f > ; < ; ? ? ; ¼; 6¼g; and (3) ri ?Ah rj. 据此，我们可以为表1中的一个实体的图元画出货币约束，具体如下

c1 : 8s; t : ðs½Salary? < t½Salary?Þ?!ðs ? Salary tÞ: c2 : 8s; t : ðs½Level? < t½Level?Þ?!!ðs ? Level tÞ: c3 : 8s; t : ðs½Comp? ¼ t½Comp? ^ s½Title? ¼ E ^ t½Title? ¼ MEÞ ?! ðs ? Title tÞ: c4 : 8s; t : ðs½Comp? ¼ t½Comp? ^ s½Title? ¼ RA ^ t½Title? ¼ RÞ ? ! ðs ?Title tÞ。一般来说，货币约束被定义在属性上。这有助于初步确定图元之间的货币顺序。我们在定义2中简要介绍如何在图元上推导出图元的货币顺序。

定义2（图元的货币秩序）。让ri; rj是两个图元，在D中ri½eID? ¼ rj½eID?的两个图元。如果满足以下条件，则rj比ri更有活力，用ri ? rj表示：i）存在A0 ? attrðRÞ, 8Ak 2 A0, t h a t ri ? Ak rj, a n d）6 9Ah 2 attrðRÞnA0, rj ? Ah ri。从CCs推导出的图元的货币顺序满足非反身性、反对称性和反证性。它是确定数据集中所有图元的货币的基础。因此，我们提出了图元的货币图G（见定义5），并通过第3节中的反式闭合发现来构造货币顺序图Gc（见定义6）。条件函数依赖（CFD）已经被开发出来，用于检测和解决数据集中的不一致问题[10]。我们在我们的方法中采用CFDs来提高数据的一致性，如定义3中所讨论的。

定义3. R上的CFD被定义为RðX ! Y; TpÞ, 其中i) X ! Y是一个标准的FD，ii) Tp是一个具有X和Y的属性的表。这里，X ? attrðRÞ; Y ? attrðRÞ，对于每个属性A 2 X [ Y和每个模式元组tp 2 Tp，tp½A? 是domðAÞ中的一个常数，或者是一个未命名的变量"\_"，从domðAÞ中取值。请注意，一个Tp中的每个模式元组tp表示一个约束条件。一个CFD ðX ! Y; tpÞ，如果tp只由常数组成，即对于所有的属性A 2 X [ Y，tp½A? 是一个常数值，则被认为是一个常数CFD。在本文中，我们重点关注这种恒定CFD，以检测同一元组的属性的不一致性。我们把'表示为一个带有模式元组tp的约束。例如，给定的CFD：ð½Comp?; ½Group? ! ½City?Þ 对于表1，我们有三个CFD的语义实例如下。



我们在本文中的工作重点是清理有多种缺陷的数据库。1）缺少时间戳的图元，2）缺少某些属性值的图元，以及3）在违反CFD方面不一致的图元。为了修复这些缺陷，我们应用上述的CCs和CFDs。此外，在归入缺失值时，我们假设对于每个属性A，有一个置信度阈值sðAÞ，归入的值必须达到这个阈值才能被接受。由于在信息不完全的情况下，归因结果可能不是很准确，所以可以人为地设置sðAÞ，以避免无意义的归因结果。我们考虑到这三个缺陷，数据集D是低质量的。定义4概述了我们的问题定义。

定义4（问题定义）。给定一个低质量的数据集D，包括一个关系模式R的大量实例，数据质量规则包括一组CC的F和一组CFD的S。对D的数据清理包括三个任务。1）将部分货币顺序扩展到D中的所有图元，并为每个图元计算一个可比较的货币价值。货币值被视为一种相对的时间戳，必须与CC的原始顺序一致；2）检测数据集D对CFD集的满足程度，用模式元组tp修复含有不一致属性值的元组，使D满足F；3）用保证归入的属性值的置信度不小于给定阈值sðAÞ的方案来归入元组中缺失的属性值。

2.2 框架概述

众所周知，完整性和一致性是侧重于图元值的质量测量的指标，而货币总是侧重于整个数据集中图元的时间顺序[9]。因此，我们沿着货币顺序处理一致性和完整性的修复。我们开发Imp3C是为了达到两个目的。根据已知的数据质量规则，即CCs和CFDs，i）Imp3C中的每个修复操作不会导致任何违反3C问题的新错误，ii）在处理Imp3C后，3C上存在的错误很少。Imp3C实现了对货币、一致性和完整性的整体数据修复，主要有四个步骤，如图1所示。 1）我们首先为具有货币约束的图元构建货币顺序图，并检测货币图中的冲突。(第3节)。(2) 我们确定从CC中提取的图元中的货币顺序，它作为图元之间关于货币的直接和明确的度量而得到。我们通过迭代更新有效边来确定图元的货币价值，并发现货币图中最长的货币顺序链。我们还测量元组对的货币顺序的可靠性，以便更合理地呈现元组中的货币顺序。这一步将在第4节讨论。 (3) 我们在不完全性之前做不一致修复。对于一致性修复问题，我们首先输入一致性约束（如CFD），并从原始数据集中提取潜在的一致性模式，以捕获未发现的一致性表。在一致性模式确定后，我们定义了一个度量CC-DIST来衡量dirty脏数据和干净的数据。我们不仅根据一致性模式来修复错误，而且还考虑到全局货币秩序，即用最接近当前时间点的适当数值来修复脏数据。这个过程将在第5节中报告。 (4) 在检测元组对CFD的满足程度时，我们将元组与CFD模式元组匹配时的缺失值作为一种违反一致性的情况。一些缺失的值可以通过给定的CFD进行修复。这对后面的归并过程的训练精度是有利的。此外，两步归因法在整个数据清理过程中节省了时间。在最后一步，我们用贝叶斯策略修复剩余的不完整值。我们将货币顺序作为一个加权特征，如果填充概率不低于A上的置信度sðAÞ，则训练完整的记录来填充缺失的数值，详细步骤见第6节。修复后的部分将在3C上引起少量新的违规。一方面，综合货币排序方法确保Imp3C提供的是具有最接近货币度量的清洁值。另一方面，根据已知的一致性约束，由S修复的缺失值不会导致一致性方面的新违规。此外，Imp3C的计算时间成本也有所下降。这是因为Imp3C在一致性和完整性问题之间采用合理的修复顺序，防止了一些不必要的修复。

3 货币顺序图的构建

我们的目标是用图来确定图元的货币顺序。本节将讨论货币图的构造。我们在定义5中提出了货币图G，根据这个定义，任何两个节点之间都有一条时间上相关的边。为了简明扼要，我们根据定义6，使用反式还原法，从G中得到货币序图Gc。定义5（货币图）：rðEÞ ¼ fr1; . . ; rng表示R的n个图元。一个有向无环图G¼ðV; EÞ是D中E的货币图，如果它满足：i) 8vi 2 V; ri 2 rðEÞ; vi ¼ ri, and ii) 9eðm; kÞ 2 E iff rm ? rk由定义2推导而得。定义6（货币秩序图）。对于一个给定的货币图G ¼ ðV; EÞ，其子图Gc ¼ ðV; ETRÞ是G的反式还原图，iff i) G和Gc有相同的反式闭合，和ii) 8G 0 c ? Gc，G 0 c的反式闭合与G的反式闭合不同。这样的Gc被称为货币秩序图。有了定义5和定义6，我们让Mike和Helen是表1中的两个实体E1和E2，并在例2中为它们构造货币顺序图。

1. 我们从第2.1节的CC推导出，图2a中的r2 ? "工资 "r3 ? "工资 "r4 ? "工资 "r5, r6 ? "职称 "r7。同样地，与



在构建货币图时，我们需要解决货币约束中的冲突，以实现准确无误的货币秩序。因此，CC中的冲突可以通过发现Gc中的循环来确定。冲突可能来自模糊的货币约束或某些图元中明确的货币问题。如果没有可靠的外部知识，这些冲突就不能被解决。由于冲突只发生在数据的一小部分，它们将被退回进行人工处理（例如，由领域专家进行修复或分配众包任务[12]）。