信息系统模拟数据生成研究综述

曹建军 刁兴春 张 慧 谭明超 邓 波
(总参第63研究所 南京 210007)

摘 要 信息系统模拟数据生成是提供信息系统试验、试用和演练中所需数据的重要途径。通过与软件测试数据生成、样本数据扩充、虚拟现实相关研究领域比较,讨论了信息系统模拟数据生成的研究定位;归纳了信息系统模拟数据生成的研究内容;提出了具有数据层、中间层和生成层三层结构的典型信息系统模拟数据生成系统结构框架;最后对信息系统模拟数据生成的研究方向进行了展望。

关键词 信息系统,模拟数据生成,软件测试数据,样本数据,虚拟现实
中图法分类号 TP31 **文献标识码** A

Simulated Data Generation for Information System: A Survey

CAO Jian-jun DIAO Xing-chun ZHANG Hui TAN Ming-chao DENG Bo
(The 63rd Research Institute of the PLA General Staff Headquarters, Nanjing 210007, China)

Abstract Simulated data generation for information system is an important way for providing data, which is needed in experiments, trials or exercises of information system. In this paper, the simulated data generation for information system was compared with other correlative fields, such as software test data generation, sample data expansion and virtual reality, and its research orientation was discussed. Research contents of the simulated data generation for information system were summarized. A typical system architecture framework for the simulated data generation for information system was proposed. The framework was divided into three layers. They are data layer, intermediate layer and generation layer. At last, the future research directions of the simulated data generation for information system were analyzed.

Keywords Information system, Simulated data generation, Software test data, Sample data, Virtual reality

信息系统在试验、试用、演练中需要数据支持,但常常不宜使用真实数据。典型原因如下:一是出于安全性考虑,不允许使用真实数据[1,2]。为了解决这一问题,在某些场合常采用一些简单变换对真实数据伪装,但伪装数据通过逆变换可以复原为真实数据,仍然存在安全隐患,同时,有些伪装数据会降低数据的合理性,影响应用效果。二是出于时间或成本要求[1,2],难以使用真实数据。真实数据需要按照一定的流程获取,并要有相应人员、装备保障,还要有最低限度的时间保证,在有时间限制或投入受限的情况下,难以获取满足要求的真实数据。三是不同任务有具体的数据需求,真实数据不能满足应用。信息系统的试验、试用、演练任务目标明确,也需要与目标相适应的数据支持。通常,已有的真实数据并不一定立足于本次任务的数据需求所获得,在完整性(completeness)、及时性(timeliness)、可用性(availability)、相关性(relevance)、效用性(transactability)等方面与实际数据需求存在偏差。

鉴于以上原因,信息系统模拟数据生成的现实需求日益迫切,应采用合理模型与算法,用计算机模拟的方式快速生成模拟数据,以满足信息系统试验、试用、演练的实际需要。本文对信息系统模拟数据生成所涉及的相关基本问题进行探讨,为该领域的全面深入研究打下基础。

1 信息系统模拟数据生成的研究定位

1.1 相关研究领域

在信息技术的不断发展过程中,产生了不同的数据生成需求,也相应催生了新的研究领域。

1.1.1 软件测试数据生成

一般意义上,软件测试是为了发现错误而执行程序的过程。顾名思义,软件测试数据是用于软件测试的数据。测试数据生成可以被理解为一个抽样过程,即根据相应的测试覆盖标准,采用一定的方法,在测试数据全集中进行抽样,选取出一批错误敏感的测试数据,使它们具有满足要求的发现软件错误的可能性[3,4]。

当前,软件已经渗透到各行各业,深刻影响着工业、农业、国防、管理以及人们的日常生活,其质量受重视的程度也越来越高,而软件测试作为软件质量保证的重要手段伴随着软件

本文受中国博士后科学基金特别资助项目(201003797),中国博士后科学基金项目(20090461425),江苏省博士后科研资助计划(0901014B)和解放军理工大学预先研究基金项目(0901014B)资助。

曹建军(1975—),男,博士后,CCF会员,主要研究方向为信息质量、进化计算等,E-mail:jianjuncao@yeah.net;刁兴春(1964—),男,研究员,博士生导师,主要研究方向为数据工程等;张 慧(1982—),男,工程师,主要研究方向为数据工程等;谭明超(1979—),男,博士生,主要研究方向为信息质量等;邓 波(1977—),男,工程师,主要研究方向为数据工程等。

本身的发展。据统计，软件测试时间会占到总开发时间的40%，一些可靠性要求非常高的软件，其测试时间甚至占到开发周期的60%；对某些特殊软件，测试费用甚至高达其他软件费用的3～5倍[4,5]。一个有效的测试数据设计方法可以生成高质量的测试数据，并尽可能降低测试数据集的规模，从而提高测试效率，缩短软件开发周期和降低开发成本。因此，软件测试数据生成当前仍是十分活跃的数据生成研究领域[3-12]。

1.1.2 样本数据扩充

在数据挖掘、可靠性评估等研究中，要求样本量足够大，而获取足够量的真实样本数据常常很困难(例如，现实中不允许做过多破坏性试验，如关键部件的损毁性试验，某些装备的可靠性试验等)。传统统计学所研究的是一种渐进理论，由此提出的各种方法只有当样本数目趋向于无穷大时，其性能才有理论上的保证。即使是近年发展起来的针对小样本的支持向量机理论，仍需要一定量的训练数据进行训练[13]。

样本数据扩充是为了解决样本数据量不足的问题，借助合适算法模型，依据少量样本或建立数学模型，生成满足要求的一定数量的样本数据。当前，样本数据扩充所涉及的数据以数量数据(quantitative data)为主。

样本数据扩充在手势识别[13]、人脸识别[14]、手写体识别[15]、文本分类[16]、可靠性分析[17]等领域得到了研究与应用，文献[2]应用遗传算法研究了数据挖掘中的样本数据扩充方法。

1.1.3 虚拟现实

虚拟现实(virtual reality)是以计算机技术为核心，结合相关科学技术，生成与一定范围真实环境在视、听、触感等方面高度近似的数字化环境，用户借助必要的设备与数字化环境中的对象进行交互作用、相互影响，用以产生亲临对应真实环境的感受和体验[18,19]。

虚拟现实产品使参与者可直接参与和探索虚拟对象所处环境的作用和变化，仿佛置身于现实世界中，产生感知性(multi-sensory)、沉浸感(immersive)、实时性(real-time)和交互感(interactive)[19]。

要在数字空间中模拟现实世界中的对象和状态，就需要将现实世界中的对象、对象之间的关系、对象之间相互作用及发展变化所遵循的规律映射为数字空间中的各种数据表示，这一过程称为建模。建模首先涉及模型数据，从来源来说，模型数据可分为实际测量、数学生成和人工构造三大类[18]。本文所讨论的信息系统模拟数据生成可直接用于虚拟现实的模型数据生成。

虚拟现实的研究最早开始于20世纪20年代末，我国于20世纪70年代初开始进入该领域[18]。随着各行各业对虚拟现实技术的需求，该领域的研究实践活动日益活跃。虚拟现实技术已在训练模拟、空间技术、计算机可视化、医学、虚拟战场等领域得到了广泛应用[18-21]。当前的研究热点主要集中于数据获取与处理、绘制、显示、增强现实与定位和人机交互设备与系统等方面[20,21]。2010年7月《中国计算机学会通讯》对虚拟现实进行了专题讨论。

1.1.4 信息系统模拟数据生成

依据不同的开发应用阶段，信息系统中的数据可以分成测试数据、模拟数据和真实数据。测试数据即为软件测试数据，真实数据是系统真正投入使用所需的数据以及运转过程中产生的数据。模拟数据又叫演练数据，指信息系统在试验、试用、演练中使用的非真实数据，这类数据对信息系统而言又具有相当的真实性。信息系统模拟数据生成是用计算机模拟的方式快速生成信息系统模拟数据。从时效性、安全性、经济性的角度，在许多场景下都需要生成信息系统模拟数据。

当前，信息系统中使用的主要是关系型数据。文献[22]研究了审计数据的模拟数据生成，开发了审计数据模拟数据生成系统，系统可以生成不同规模、含有不同长度要求的相似重复数据或不符合业务规则的数据；系统主要借助开源网站提供的数据为源数据，通过调用源数据来生成所需的数据，没有考虑关系型数据的特点，所生成的数据并没有真实的领域背景，适用性有限。文献[1]研究了人员档案的模拟数据生成，设计实现了一个人员档案模拟数据生成系统，用以生成人员的姓名、性别、民族、出生年月、籍贯、政治面貌、学历、学位等属性值；系统利用随机数函数，随机选择给定域的值生成各属性值，没有考虑各属性值的分布及属性值间的依赖关系。

DB Data Generator 和 SQL Data Generator 是分别由 Datanamic 和 Red Gate 公司开发的两个关系型数据生成工具，但当前二者功能有限，仍然仅支持单属性值生成，生成过程没有考虑属性间的关系。

1.2 信息系统模拟数据生成的研究范围

信息系统模拟数据生成的研究范围如图1所示。

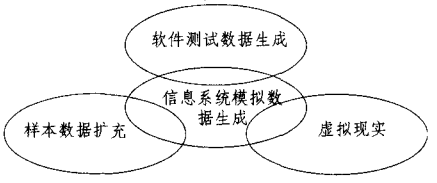

图1 信息系统模拟数据生成的研究范围

如图1所示，信息系统模拟数据生成的研究范围和软件测试数据生成、样本数据扩充及虚拟现实有明显重叠，同时又有不同于其他三者的独立研究范围。

1.3 信息系统模拟数据生成的特点

尽管信息系统模拟数据生成的研究范围与软件测试数据生成、样本数据扩充和虚拟现实没有明确界限，并且在同一个信息系统的开发应用周期内，可能前3种数据生成需求并存，但是，以上4者具有明显区别：

(1)任务需求不同。软件测试数据生成可以被理解为一个抽样过程，即根据相应的测试覆盖标准，采用一定的方法，在测试数据全集中进行抽样；样本数据扩充是为了解决样本数据量小的问题，根据少量样本获取大量样本的过程；而信息系统模拟数据生成解决的是在不方便或不可能使用真实数据时，生成信息系统正常运行所需的数据；虚拟现实的数据来源更加多样，信息系统模拟数据生成可以作为其获得数据的具体手段。

(2)生成过程不同。软件测试数据生成可以看成在完整的数据空间内搜索测试数据集的过程，可以用组合优化问题对其进行建模；样本数据扩充是一个由少生多的过程；信息系统模拟数据生成可以包括由少生多的处理，但更重要的是通过对关系型数据库中依赖关系、规则的描述，从无到有的数据生成过程。

(3)处理数据不同。软件测试数据生成是从已知的数据集(可以看成一个二维矩阵,矩阵的列可以是软件的配置参数、内部事件、外部输入等)中进行抽样;样本数据扩充以数量数据为主,这些数据携带了所描述对象某一状态特征信息;由于现实世界的对象千变万化,虚拟现实涉及的数据复杂多样,通常包括图像、视频等非结构化数据;记录是关系型模拟数据生成的基本单位,关系是关系型数据的基本特性,所以对"关系"的处理是当前信息系统模拟数据生成的关键。

(4)任务目标不同。软件测试数据生成是追求覆盖和测试集规模的最佳折衷过程,即追求在尽可能小的测试数据集中得到更好的测试结果;样本数据扩充的目标是使生成的数据既能达到数据量要求,又要确保保留初始样本数据所描述对象状态的基本特征信息,同时还应使生成的样本具有一定的多样性;虚拟现实的最终目标是实现对真实场景的模拟,而模型数据是虚拟现实建模的基础;关系型数据的生成要求生成的数据要满足指定的完整性约束条件,并且要符合指定的领域业务规则,还要求满足对数据集的特殊要求,如某单位的人员记录表可能要求与实际人员类别编制情况相吻合。

2 信息系统模拟数据生成的研究内容

信息系统模拟数据生成包括以下研究内容:

(1)数据生成需求分析。每一个具体的信息系统模拟数据生成任务,在实施数据生成之前,都要进行详细的需求分析,给数据生成提供依据。需求分析应包括充分理解用户的数据生成需求,就具体的生成数据质量与数据用户达成一致,并最终形成模拟数据生成详细实施方案。

(2)数据生成方法研究。以当前最为典型的信息系统关系型模拟数据生成为例。关系型数据不但具有严格的结构,并且各属性值之间存在各种数据内部依赖(internal dependency),数据生成需要发现并描述这些依赖关系,实现数据生成,使生成的数据保持所要求的依赖关系;关系型数据中含有多种属性类型,并且它们的定义往往与业务领域相关,因此,对于不同类型的属性,有必要分别研究相应的生成方法;一个完整的关系型数据生成任务,特别是海量数据生成任务,还要研究数据生成流程的优化问题。

(3)生成数据的评价。数据生成之后,有必要对生成数据与模拟数据生成需求的符合性、可用性等进行评价。具体评价方法分为主观定性评价和客观定量评价。主观定性评价可以请用户或有经验的专家从不同角度打分评价,客观定量评价通过计算生成数据的特征参数指标值进行评价,最后往往需要综合主观评价与客观评价给出最终的评价结论。模拟数据评价的核心问题是确定评价指标与评价(打分)标准,对定量评价还要研究不同指标的计算模型,如何将多个指标最终融合成一个科学合理的评价结论等。

(4)生成数据的私密性与安全性。所生成的关系型数据,不可能完全消除原数据集的真实信息以及一些领域信息。事实上,有时也往往要求所生成的模拟数据保留部分真实信息,以达到特定的目的,如"迷惑性"等。因此,所生成的数据仍然存在私密性与安全性问题,有必要对此进行专门研究,并且该问题与数据的符合性、可用性之间存在矛盾,在一个数据生成任务中,有时需对两个问题进行折衷考虑。

3 信息系统模拟数据生成系统的结构

一个典型的信息系统模拟数据生成系统具有如图2所示的结构。

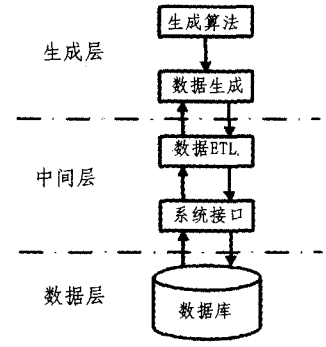

图2 信息系统模拟数据生成系统结构示意图

如图2所示,典型的信息系统模拟数据生成系统具有3层结构:数据层、中间层和生成层。

数据层:是生成数据的最终存放文件,可以是一个关系型数据库,也可以是一张简单的Excel表。对一个使用关系型数据库的信息系统模拟数据生成,该部分是一个完整的关系型数据库文件,包括数据结构、数据字典等模拟数据生成所需信息。

中间层:包括系统接口与数据ETL(Extraction, Transformation and Loading)两部分。常用的关系型数据库有Oracle、SQL、Access、Db2、Sqlserver、Sybase等,系统接口应能完成对所要求格式类型的数据库进行读写,以实现模拟数据生成系统的通用性;数据ETL主要完成对数据库的解析,并根据数据生成要求,依据参照完整性规则等,对数据库中的表进行一定转换。

生成层:包括数据生成与生成算法两部分。生成算法是一个算法集合,这些算法根据要生成的属性值类型或多个属性值之间的依赖关系进行设计,对算法的管理是生成系统的重要功能,生成算法应包括一个完善的算法管理机制;数据生成是根据要生成数据的表(一般是经过变换的逻辑表),通过选择与单个属性或包括多个属性的属性组对应的算法进行数据生成的过程,将所生成的数据存放在逻辑表中,数据库包括多个表,但数据生成以一个逻辑表为处理单位,数据生成完成后,通过数据ETL经系统接口加载进数据库。

另外,为了增加关系型数据生成系统操作的友好程度,系统应该能够对数据库信息、数据生成过程等进行良好展现。

结束语 相对于软件测试数据生成、样本数据扩充和虚拟现实,信息系统模拟数据生成仍是一个崭新的数据生成研究领域,相关工作尚在起步阶段。但随着信息化的深入发展,信息系统试验、试用、演练的数据需求场景也越来越多,自然会促进信息系统模拟数据生成的更快发展。该领域将会呈现出以下发展趋势:

(1)与其他相关领域的融合发展。信息系统模拟数据生成与其他相关领域存在共性问题,借鉴其他领域的成果积累与实践经验将加速信息系统模拟数据生成的发展;

(2)对关系型数据生成的深入研究。研究重点是关系的描述及其在模拟数据生成中的应用,该研究过程将用到当前

(下转第338页)