基于多数据源可动态配置的用户信息引接模型

丁文超[1] 伍 荣[1] 杨少鹏[1] 万思思[1] 周 猛[1]
DING Wenchao WU Rong YANG Shaopeng WAN Sisi ZHOU Meng

摘 要 伴随大数据、云计算技术的逐渐成熟和应用，以云模式来共享用户数据成为新的发展方向。传统的业务系统中，用户信息系统独立分散在不同的地方，系统之间的用户信息并没有互联互通。针对各个系统中存储的用户信息数据项参差不齐，从各用户系统获取到的用户信息汇聚成统一的用户数据进行统一管理和安全保护成为比较大的难题，文章提出基于多数据源的可动态配置的数据引接模型。该引接模型实现用户信息的引接并统一管理，能够形成统一的用户数据格式，实现数据高效共享，同时提供了高效的数据模型匹配方法，能够完成不同数据源的用户信息转化，具有实际的指导作用和广阔的使用前景。

关键词 多数据源；信息引接；动态配置；用户信息；数据模型

doi：10.3969/j.issn.1672-9528.2025.08.039

0 引言

传统业务系统中，用户信息系统独立分散在不同的地方，系统之间的用户信息并没有互联互通，缺乏统一的管理手段，用户信息的安全保护措施也参差不齐，缺乏统一的保护，同一用户在不同身份管理系统注册时，由于个体状态变化（如职位、办公地点变更）、操作失误等原因，导致相同属性的内容不同，并且会因业务系统设计要求导致相同属性的格式有差异[1-3]。因此，需要在用户信息引接过程中考虑内容冲突和统一数据格式问题。同时，各个系统中存储的用户信息数据项也参差不齐，如何将来自各用户系统的用户信息汇聚成统一的用户数据进行统一管理和安全保护成为很大的难题。

伴随大数据、云计算技术的逐渐成熟和应用，以云模式来共享用户数据成为新的发展方向[4-5]。在实际的应用系统中，用户数据来自多个系统，需要从不同的用户信息系统中引接用户数据，并利用数据模型映射机制将用户数据转换为标准化格式，进而对转换后的用户数据进行集中管理，降低各系统间信息的冗余，减轻用户信息维护的负担。

目前，针对数据融合的技术[6]大多是针对数据采用具体的算法如矩估计、最小二乘法等来进行数据的融合[7]，主要应用在大数据、人工智能领域，针对具体的应用系统中的数据去拉取并形成统一的数据格式的模型很少，基于多数据源的可动态配置的数据引接模型针对实际应用场景提出的数据引接方案，更具有实际的指导意义和广阔的应用价值。

为了克服现有技术的上述缺点，提出了一种基于多数据源的可动态配置的用户信息引接模型，旨在解决多数据源的用户信息引接问题。

1 技术方案

1.1 总体工作流程

基于多数据源的可动态配置的用户信息引接模型的总体工作流程如图 1 所示，包含以下过程：

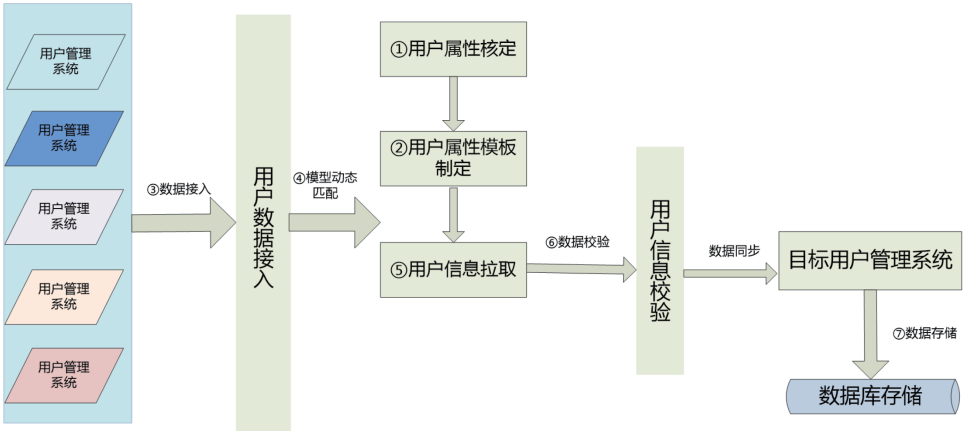

图 1 总体工作流程图

（1）用户属性核定：获取数据源用户模型字段 $D_1, D_2, \cdots, D_m$。

1. 中国电子科技集团公司第三十研究所 四川成都 610041

（2）用户属性模板制定：各个数据源的用户信息所存储的属性从名称、数量、关键字段设置上不尽相同，会造成用户信息源数据到目标数据直接引接过程中的兼容性问题。因此完成根据需要引接的数据源的用户字段编写匹配模型，匹配当前统一用户管理系统中的用户模型字段 $F_1, F_2, \cdots, F_n$。

（3）数据接入（包括数据清洗）：根据数据源用户管理系统提供的数据获取方式，获取用户管理系统中的用户数据；从各个数据源抽取的数据质量并不能得到保证，会临时存放在中间数据表中，通过初步的数据分析进行数据清洗处理，主要包括对缺失值、逻辑错误、格式有误、无关属性的脏数据清洗 [8-9]。

由于数据关键字段内容的不完整，需要进行缺失值清洗，有三种方法解决：一是直接删除，根据数据可用程度决定是否删除掉缺失的记录；二是采用均值填充法，用该列数据的平均值来填补缺失字段的内容；三是高频填充，使用当前列出现频率最高的数据对该字段内容进行填充。

格式内容清洗主要包括：对数据中出现的空行进行删除；对列数据的单位进行统一，如身高信息的单位，有的数据源定义为“厘米”，而有的则定义为“米”；对列属性的合并，如有些源数据把“姓名”属性分为“姓”和“名”两个属性；对非 ASCII 字符进行转换。

逻辑错误清洗主要指数据之间的冲突处理以及重复数据删除。

在无关属性清洗中，需要对非目标字段进行识别和剔除，以免影响后续模型匹配的效率。

（4）模型动态匹配：根据当前数据源的配置信息动态加载数据匹配模型。

（5）用户信息拉取：将数据源用户管理系统中的数据接入目标用户管理系统中。

（6）数据校验：在向目标数据库导入数据前，需要完成数据校验工作。针对每一条引接的用户信息进行目标数据库的全表扫描，将重复或内容冲突的数据，根据事先定义的数据来源级别和可信程度形成优先级数据，完成高优先级数据对低优先级数据的替换。

（7）数据同步：将校验后的数据保存到目标用户身份管理系统中。

1.2 数据引接模型实现

数据引接模型实现包括动态配置设计、数据引接模型实现、数据匹配方法，其中，动态配置设计针对多种数据源环境下，动态配置新的用户信息数据源引接模型，实现用户信息的引接并统一管理；数据引接模型实现通过数据模型映射解决不同数据源的用户数据转化，形成统一的用户数据格式，实现数据高效共享；数据匹配方法提供了高效的数据模型匹配方法，解决不同数据源的用户信息转化功能。

1.2.1 动态配置设计

在一个设计良好的信息系统中，上层模块（通常负责业务逻辑）不应直接依赖下层模块（如网络通信、数据库访问等），而是通过抽象接口调用下层模块功能。上层模块依赖抽象接口，而下层模块负责具体的功能实现。这种“依赖倒置”的设计模式便于系统更好地适应后续的变化和功能扩展。通过抽象接口解耦上层模块与下层模块之间的依赖关系，上层模块无需关注具体实现细节，从而使系统设计更加灵活、易于扩展和维护。

在设计多源数据接入模型时，对外部数据源的接口调用由系统的下层模块负责，而业务流程中由上层模块负责数据的处理。通过依赖倒置原则 [10]，上层模块定义一个数据访问对象接口，具体的接口实现由下层模块提供，从而实现实际的数据接入操作。针对不同的外部数据源访问方式，可以定义不同的底层实现类，当需要切换数据访问方式时，只需通过配置文件更改接口实现类，而高层的业务处理模块无需任何修改，从而实现可动态配置的目标。该设计思路不但提高了信息系统的灵活性，而且增强了模块的可维护性与扩展能力。

1.2.2 数据引接模型实现

如图 2 所示，在目标用户管理系统中，通过配置文件配置不同数据源的用户模型及数据源数据拉取服务具体实现，利用动态反射机制来实现用户数据模型动态匹配和数据获取 [11]，从而完成数据源用户管理系统中的用户数据到目标用户管理系统中的数据引接。

基于多数据源的用户数据动态匹配工作流程如下：

（1）数据获取接口调用：目标用户管理系统调用数据源获取服务拉取数据。

（2）数据模型配置解析：数据源获取服务内部通过工

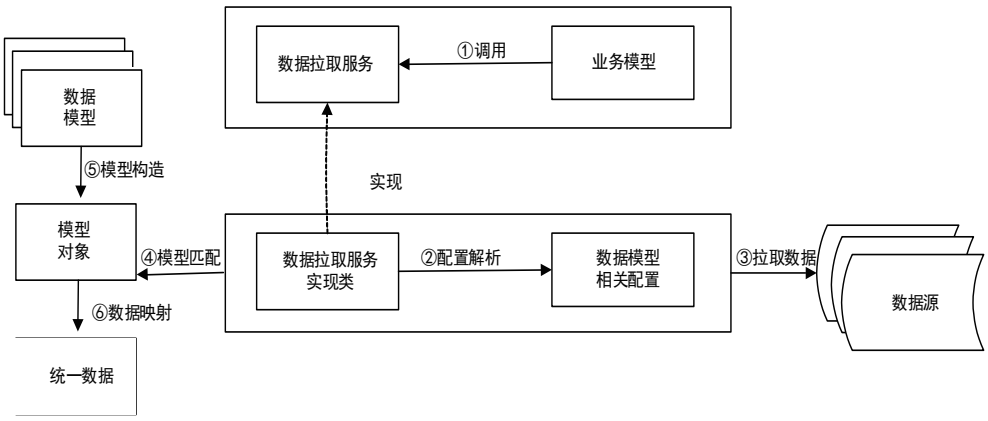

图 2 多数据源引接实现图

厂化的方式解析对应数据源的配置文件，获取具体的数据源获取服务实现类。

（3）拉取数据：根据解析配置文件得到的具体的数据拉取实现，获取外部用户管理系统中的用户数据。

（4）模型匹配：根据配置文件中的数据匹配模型，加载对应的模型匹配结构。

（5）模型构造：根据加载得到的模型结构，构造具体的模型对象。

（6）数据映射：根据模型对象中的数据映射规则，将从数据源用户管理系统中拉取到的用户数据映射到目标用户管理系统中的用户模型中。

（7）数据返回：将映射后的统一的用户数据返回给目标用户管理系统调用服务，通过进一步处理之后，存储到目标用户身份管理系统中的数据库中。

1.2.3 数据匹配方法

为了形成最终的统一的用户数据，从各数据源接入的数据，其数据项字段设为 $D_1, D_2, \cdots, D_m$，目标用户管理系统中的用户模型字段 $F_1, F_2, \cdots, F_n$，其中（$1 \le m \le n$），根据数据中集合的映射关系，（$F_1, F_2, \cdots, F_n$）=f（$D_1, D_2, \cdots, D_m$），由于 $1 \le m \le n$，所以该映射为一个多对多的映射模型。在实际使用该映射模型进行映射时，如果 $m < n$，则需要对多余的数据项进行默认值 $V_{m+1}, V_{m+2}, \cdots, V_n$ 填充，数据字段之间采用一一映射的方式来构建最终的用户数据。

最终采用的模型映射方式如下：

当 $1 \le m = n$ 时，$(F_1, F_2, \cdots, F_m) = f(D_1, D_2, \cdots, D_m)$；

当 $1 \le m < n$ 时，$(F_1, F_2, \cdots, F_n) = f(D_1, D_2, \cdots, D_m, V_{m+1}, V_{m+2}, \cdots, V_n)$。

2 具体实现

假设数据源为某银行用户管理系统，需要从该用户管理系统中引接用户数据，当用户数据源接入到用户身份管理系统时，需要对每组数据源的数据进行初步清洗，同时根据属性比较建立匹配的数据引接模型，引接过程中对数据内容进行边界安全、内容冲突等相关校验。在系统设计时可以采用基于抽象工厂构建模式的框架模型，支持不同数据源的动态扩展，方便用户身份管理系统从不同的系统中引接用户数据，实现用户数据和其他系统的互联互通。主要通过以下几步实现数据引接：

（1）核定该数据源用户模型字段，包括：员工姓名（d_userName）、身份证号（d_idNo）、员工 ID（d_uid）、员工职务（d_duty）、员工性别（d_sex）、员工职级（d_dutyLevel）、员工月收入（d_salary）。

（2）假设需要引接的用户数据有员工姓名（d_userName）、身份证号（d_idNo）、员工 ID（d_uid）、员工职务（d_duty）、员工性别（d_sex）、员工职级（d_dutyLevel）。目标用户管理系统中用户属性字段有姓名（userName）、身份证号（idNo）、用户 ID（uid）、职务（duty）、性别（sex）、职级（level）、平台 ID（pid）。根据该数据源用户管理系统中用户属性字段与目标用户管理系统中用户属性字段的映射关系配置匹配模型：

(userName, idNo, uid, duty, sex, level, pid) = *f*(d_userName, d_idNo,d_uid, d_duty, d_sex, d_dutyLevel, d_idNo)，其中平台 ID 使用身份证号作为默认值来填充。

（3）配置引接数据源目标用户管理系统中的用户数据使用的服务。

（4）通过统一的数据获取服务从数据源中提取用户数据。

（5）在目标用户管理系统调用数据拉取服务执行过程中，会进行模型的匹配以及数据的映射，拉取过来的用户数据会在校验完成之后进行数据的同步，最终引接到目标用户身份管理系统数据库。

3 结语

本文提出的基于多数据源的可动态配置的用户信息引接模型，能够解决多数据源的用户信息引接问题。提出的数据匹配方法适用于各种需要进行数据共享的业务场景，解决了不同数据源的用户信息转化，具有较广阔的应用价值；针对多种数据源环境下，提出的动态配置设计，动态配置新的用户信息数据源引接模型，实现用户信息的引接并统一管理，为实际应用场景中的数据共享实现提供了自动化、智能化手段，减少了人工干预过程；针对不同数据源的用户数据转化问题，提出的数据模型匹配流程实现，能够准确地将不同数据源的用户数据转换成统一格式的用户数据，特别是针对需要从众多数据源系统中拉取用户数据的应用场景，此流程的实现能够有效进行数据的整合，构建统一的用户数据格式，确保多源数据高效流通和共享。

参考文献：

[1] 高宏，谢丰．数字图书馆推广工程统一用户管理平台的设计与实现 [J]. 图书馆学研究，2017 (15):12-17.

[2] 王尚平．基于区块链的用户身份管理协议及其在物联网中的应用 [D]. 陕西：西安理工大学，2020.

[3] 张淑娥，田成伟，李保罡．基于区块链技术的身份认证研究综述 [J]. 计算机科学，2023,50(5):329-347.

[4] 卢忠媛．云计算环境下用户身份管理及分级授权机制研究 [D]. 湖北：华中科技大学，2013.

[5] 孙赢．云计算中的虚拟身份认证技术研究 [J]. 科技通报，2013, 29(2):94-96.

[6] ZHU X X, TU Y, GUO B X, et al. Research on unified user data model based on multidimensional electronic channels of internet marketing services[EB/OL].[2024-06-10].https://pdfs.semanticscholar.org/c562/e97a9b69381239018a86855dad0bea218b11.pdf.

融合动态权重的以太坊钓鱼检测图神经网络模型

暴琪璐 [1]
BAO Qilu

摘 要 随着区块链技术的广泛应用，以太坊平台上的钓鱼诈骗因其高隐藏性和破坏性成为重大安全威胁。现有检测方法多依赖静态交易特征或局部网络结构，难以捕捉动态交易模式与全局拓扑关联。文章提出一种融合动态权重的以太坊钓鱼检测图神经网络模型（Dyn-GNN）。通过设计动态权重机制，模型能自适应学习交易网络的时序演化规律，捕捉钓鱼地址的异常交互模式。实验表明，Dyn-GNN 在真实以太坊数据集上的检测准确率达 91.2%，相比基线模型提高了正确性和检测性能。

关键词 以太坊；钓鱼检测方法；动态权重方法；图特征方法

doi：10.3969/j.issn.1672-9528.2025.08.040

0 引言

区块链技术的去中心化与匿名性的特性使其成为金融创新的重要载体，但同时也催生了大量钓鱼诈骗行为 [1]。根据 Chainalysis《2023 年加密货币犯罪报告》，2022 年以太坊钓鱼诈骗造成的经济损失达 317 万美元，占 DeFi 领域总损失的 43%，且攻击生命周期中位数为 18 h，呈现显著的短时爆发特性。攻击者常通过伪造高收益智能合约地址诱导用户转账，其攻击模式动态多变且缺乏固定特征 [2]。传统的基于规则或静态特征的检测方法面临显著调整。

早期研究集中于交易特征提取：Chen 等人 [3] 将交易网络抽象为图结构，构建 8 维交易特征，结合 LightGBM 实现分类。Farrugia 等人 [4] 进一步构建 42 维特征集，并识别关键指标。Ibrahim 团队 [5] 通过特征筛选实现 6 维随机森林分类（准确率 85%）。然而，这类方法依赖专家经验设计特征，难以捕捉交易间的非线性关联与时序依赖。

部分学者通过图嵌入技术来识别钓鱼节点：经典方法 DeepWalk[6]，Node2vec[7] 等算法自动提取节点特征，针对交易特异性，Wu 等人 [8] 提出 Tran2vec 算法，将有偏游走策略与交易金额、时间戳信息结合增强了异常交易的敏感性。尽管这些方法降低了特征工程成本，但其游走策略的随机性可能导致关键路径丢失，且对动态交易的时序建模能力不足。

现在大多研究是基于图神经网络检测的方法，例如 MCGC 模型 [9] 采用多通道架构聚合多层次交易模式。Li 等人 [10] 设计动态子图采样方法，优化计算效率。然而，传统 GNN 在处理以太坊交易时存在两大瓶颈：一是固定聚合权重忽略交易金额的时效性；二是全局拓扑特征与局部子图模式的融合不足。

1. 西安石油大学 陕西西安 710065

[7] 仇昌荣，姜岳道．基于最小二乘法的生产配置参数建模系统设计 [J]. 现代电子技术，2021,44(4):83-87.
[8] 姚海龙，王彩芬，许钦百．等．基于回归模型的采集数据清洗技术 [J]. 电光与控制 ,2022,29(4):117-120.
[9] CICHY C, RASS S．An overview of data quality frameworks[J].IEEE access, 2019,7:24634-24648.
[10] 姚海龙，王彩芬，许钦百，等．敏捷设计原则与设计模式的编程实践：单一职责原则与依赖倒置原则 [J]. 计算机应用，2011,31(2):149-152.
[11] 丁春玲，路志强，彭伟．Java 反射机制在数据持久层轻量级 ORM 框架中的应用研究 [J]. 西安文理学院学报（自然科学版），2017,20(1):39-42.

【作者简介】

丁文超（1991—），男，河南商丘人，硕士，高级工程师，研究方向：网络与信息安全。

伍荣（1979—），男，四川南充人，硕士，高级工程师，研究方向：网络与信息安全。

杨少鹏（1986—），男，山东烟台人，硕士，高级工程师，研究方向：网络与信息安全。

万思思（1992—），女，四川成都人，硕士，工程师，研究方向：网络与信息安全。

周猛（1992—），男，河南商丘人，硕士，工程师，研究方向：网络与信息安全。

（收稿日期：2025-04-24 修回日期：2025-08-07）