

基于图的异构数据集成方法研究

黄跃珍^{1,2}, 杨芬¹, 田丰³, 张承业¹, 李雨婵¹

1. 广州广电运通信息科技有限公司, 广东 广州 510663;
2. 广州数字科技集团有限公司, 广东 广州 510627;
3. 广电运通集团股份有限公司, 广东 广州 510663

摘要

企业各部门对数据施行分散管理, 烟囱式的系统建设使数据散落在异构数据库中, 异构数据给当前数据集成工作带来了系列挑战。为解决企业异构系统数据汇聚融合的问题, 提出一种基于图的端到端的数据集成框架。首先, 根据关系型数据模型的主外键关系将表和字段的实体关系构建成网络图, 将表名和字段名称分别看作图中不同类型的实体。然后, 将构建的图输入图神经网络, 经过图卷积得到图中各节点的向量表征, 基于节点向量可计算任意所需匹配的两个图的节点映射关系。完成图中表和对齐后, 再将不同字段值标准化, 即将每个单元格的值映射为标准值。最后, 将以上结果工程化为数据库可执行的查询语句, 从而实现异构数据融合。在企业内部的真实数据上进行验证, 实验结果表明, 文中所提框架能提高数据集成的开发效率, 且该模型不受业务领域限制, 具有较强的移植性。

关键词

数据集成; 数据融合; 异构数据; 模式匹配; 实体对齐; 图神经网络

中图分类号: TP3

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025002

Research on graph-based heterogeneous data integration method

HUANG Yuezhen^{1,2}, YANG Fen¹, TIAN Feng³, ZHANG Chengye¹, LI Yuchan¹

1. GRGBanking I. T. Co., Ltd., Guangzhou 510663, China
2. Guangzhou Digital Technology Group Co., Ltd., Guangzhou 510627, China
3. GRGBanking Equipment Co., Ltd., Guangzhou 510663, China

Abstract

Various departments of the enterprise implement decentralized management of data, and the chimney-style system construction causes data to be scattered in heterogeneous databases. Heterogeneous data poses a series of challenges to the current data integration work. In order to solve the problem of data aggregation and fusion of enterprise heterogeneous systems, an end-to-end data integration framework based on graph was proposed. Firstly, the table and field entity relationships were constructed into a network graph based on the primary and foreign key relationships of the relational data model. The table names and field names were regarded as different types of entities in the graph. Then, input the constructed graph into the graph neural network, and the vector representation of each node in the graph was obtained through graph convolution. Based on the node vectors, the node mapping relationship of any two graphs that need to be matched can be calculated. After aligning the tables and fields of the graph, the next step was to standardize the field values, meaning that the value of each cell was mapped to a standard value. Finally, engineer the

above results into executable query statements for the database to achieve heterogeneous data fusion. Through verification on real data within the enterprise, the experimental results show that the framework proposed in the paper can improve the development efficiency of data integration, and the model is not limited by business fields and has strong portability.

Key words
data integration, data fusion, heterogeneous data, schema matching, entity alignment, graph neural

0 引言

随着企业信息化建设的不断发展，不同的业务系统中都积累了海量的数据，这些数据存储在不同类型的关系型数据库中。由于开发团队遵循的规范和标准不同，数据库的表结构、表名、字段名及字段名的含义均存在差异，甚至同厂商不同版本系统的数据库的表结构和字段命名方式也存在较大的差异，导致各业务系统之间的数据存在异构性，无法实现统一管理和共享，产生“数据孤岛”现象^[1-2]。为解决异构数据的问题，企业需要进行数据集成，将来自不同数据源、不同格式和不同结构的数据进行整合统一^[3]，这个过程需要对存量的异构数据进行清洗、转换、映射等处理，以获取标准化的、结构统一的数据。

数据集成的核心技术是数据融合。数据融合指将异构数据源的数据进行整合、合并的过程，是解决来自不同数据源的同一实体在集成过程中产生的冲突问题，从而保证数据的正确性与一致性^[4]。总体来看，数据融合分为两类场景：一是纵向的垂直领域的数据融合，也称数据合并，通过将多个异构系统的数据合并到同一张表中，得到数据全集，例如大型企业集团企业旗下存在多个下属子企业，集团需要汇聚采集各个子企业人力系统的

员工信息数据进行数据合并来得到全集团的人员数据；二是横向的跨领域的数据融合，即将异构数据源中同一事物不同表示的实体进行识别消歧，获取该实体更多的有价值的信息，以打破“数据孤岛”，实现跨源数据之间的关联互通，例如将信用卡信息的关系型数据集和用户购买商品的支付记录数据集进行融合，以获取实体更多的扩展信息。本文重点研究的是第一种场景。

企业传统的数据融合主要包括以下步骤：

- 结合业务需求梳理业务逻辑、理解系统逻辑；
- 梳理数据库数据模型与业务的联系，理解数据流向；
- 梳理同业务不同系统表中数据项枚举值间的差异；
- 设计数据融合标准和数据开发模型；
- 编写结构化查询语句（structured query language, SQL），实现数据处理融合。

每对接一个系统都需要经过以上繁杂的梳理、采集、加工流程，数据链路非常长、对接工作量极大，极其耗费时间，这便是人们常说的数据分析80%的时间被用于数据整合清洗^[5]。此外，随着上游系统数据不断增加或进行系统改造，已编写的清洗规则可能会失效，会造成数据准确性降低，从而引发新的数据质量问题，导致下游应用数据出现异常情况。图1所示为

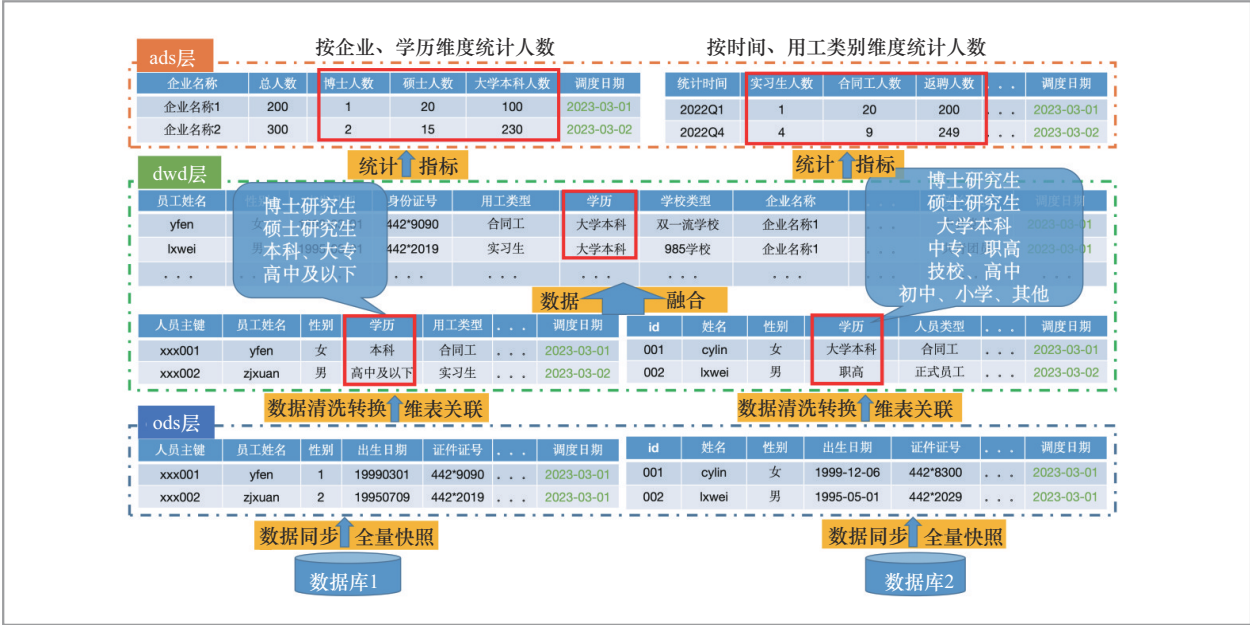


图1 传统的企业数据融合实例

不同人力资源管理系统的人员基本信息表、字段差异以及根据数据标准进行层层数据加工融合的传统的企业数据融合实例。

为解决传统的数据融合加工链路长、效率低的问题，本文提出了一种基于图神经网络的端到端的数据融合模型，将数据融合分成多个阶段，每个阶段负责不同的任务，通过多个模块的无缝连接构建数据融合一体化框架。具体来说，考虑到图结构能非常直观地表示出关系型数据集的结构信息和潜在的语义关系，笔者基于关系型数据集的结构内在属性将其编码为图结构，该结构形式包含表和字段两种节点类型。当表或字段中存在多个词时，其会被当作整体表示为一个节点，节点间通过关系型数据集中表结构的主外键和表与字段的包含关系进行连接。完成图结构的构建后，将图输入 DGMC（deep graph matching consensus）^[6]网络中，进行表和字段实体的映射对齐，将任意两个构建

好的图输入 DGMC 模型中，便可输出两个图中相似节点的匹配关系。图2第二步所示为将两个数据库表中的员工基本信息表和基本信息表对齐，同时将两个表中包含的字段也对齐。在表和字段实体映射对齐阶段，笔者还尝试用多种算法进行模式匹配效果对比。由于两个异构数据源中同一个字段的属性的值取值范围往往存在差异，因此在完成表和字段节点关系的匹配后，接下来需要对已对齐的属性值进行数据标准化转换，如考虑到预训练模型强大的语义表征，本文采用预训练模型 Debertav3^[1,7]来实现数据项值与数据标准映射的对齐。找到表和字段的映射对齐关系后，再将字段对应的值与数据标准值进行映射转换，实现数据标准化。图2第三步所示为将数据表中证件类型字段值“身份证”映射为标准值“居民身份证”。最后，将前三步的结果通过工程化方法自动化输出为可执行的 SQL 语句，从而实现数据自动融合。

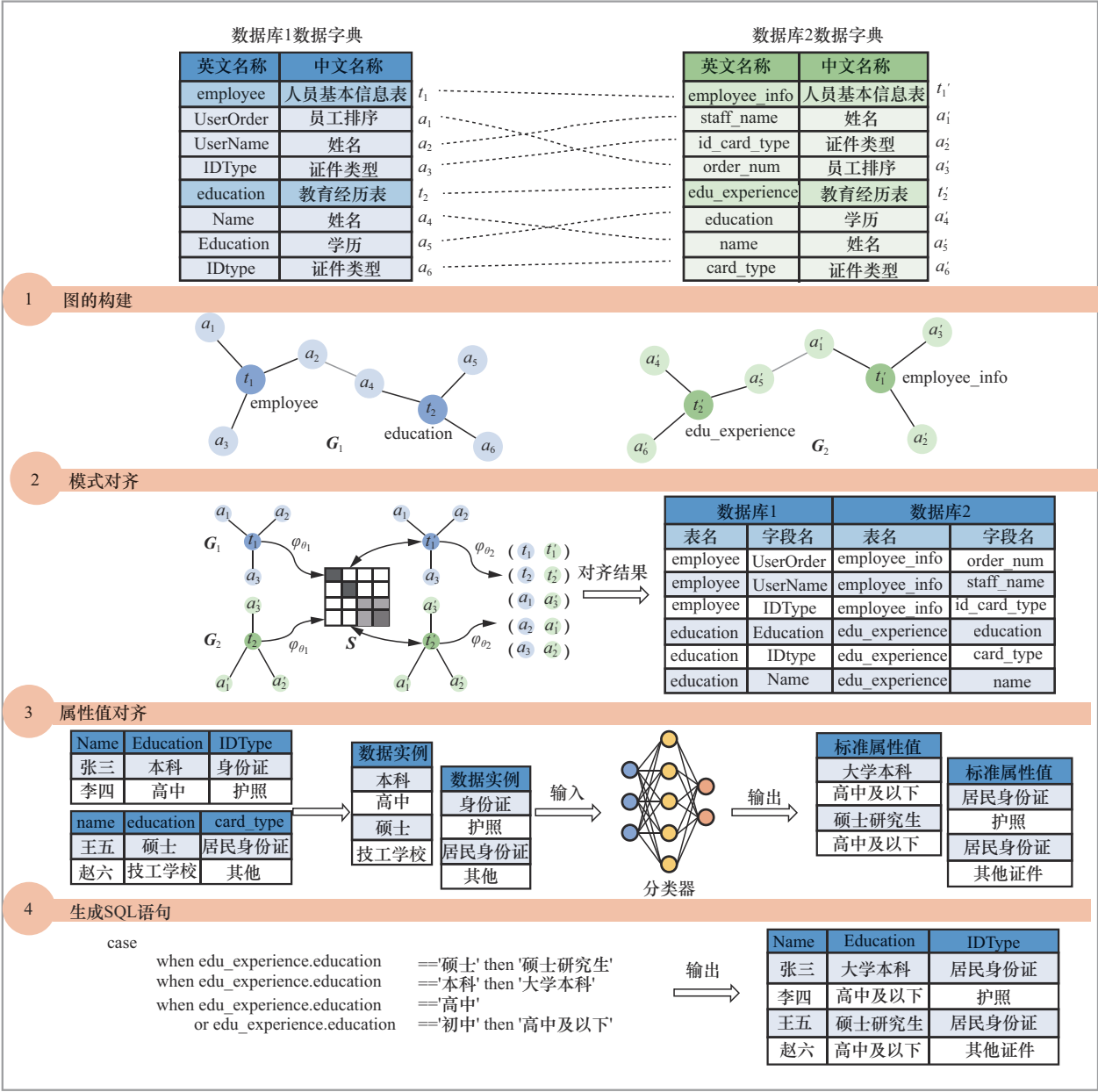


图2 数据融合模型框架

1 相关工作

数据集成包括数据清洗、关系解析、语义消歧、数据融合等技术。专家学者对以上技术的关注不断增强，其中数据融合

已成为数据集成领域的关键研究方向^[4]。最早期的数据融合方法是基于规则的方法，依靠人工定义规则来确定记录是否指代同一实体，如编辑距离计算相似度^[8]、用Jaccard相似度^[9]等来计算记录间的相似度，该方法严重依赖领域专家来定义人工规则，处理异构数据时灵活性、扩展性不

够强，因此学者开始引入机器学习的方法，如采用高斯混合模型^[10]、支持向量机（support vector machine, SVM）^[11]来解决实体对齐问题。FlexMatcher算法^[12]通过训练逻辑回归、KNN（k-nearest neighbor）等机器学习模型，将多个数据源的表结构映射到统一的表结构，当新数据源需要进行映射时，可用已训练好的机器学习模型将新数据源的每个字段分类到标准的字段名称。

考虑到数据本身包含一些特殊的内部结构信息，学者提出了基于图结构的向量表征方法，通过将数据转换为图结构来捕获不同属性之间与不同元组之间更丰富的语义关系，学习图节点和关系的特征向量，再基于向量相似度完成下游数据融合任务。其中，TransE^[13]将关系视为从头实体到尾实体的翻译，旨在将实体和关系映射到一个共享的向量空间中，以获取图结构中各节点和关系的低维特征向量。但该模型假设实体和关系在同一语义空间，难以捕捉复杂语义关系，例如复合关系和嵌套关系等。随后相继产生了TransH^[14]、TransR^[15]、TransD^[16]等改进模型。BootEA模型^[17]引入了自举策略和权重调整机制，通过迭代来提高对齐结果的准确性和鲁棒性，解决了初始对齐种子中存在的噪声和不完整性问题。该模型还利用翻译模型生成实体和关系嵌入，并借助一组对齐种子来协调嵌入空间，但该模型在处理复杂的多关系图和保持精确度方面存在一定的局限性。GCN-Align模型^[18]是一种基于图卷积网络（graph convolution network, GCN）^[19]的实体对齐模型，该模型首次提出利用图卷积网络生成结构和属性的特征向量，其利用图结构和对齐损失函数来提取实体的特征表示。RDGCN（reinforced dependency graph convolutional network）模型^[20]提出了一

种新的关系感知双图卷积网络，通过知识图和它的对偶关系之间的密切交互来整合关系信息，通过聚合实体的邻居节点信息来更新实体的特征表示。DGMG（deep graph matching consensus）模型^[6]通过图信息传播和特征聚合来捕获图结构的信息，同时结合了局部区域特征和全局图结构特征，以更准确地进行图节点匹配。Zhao等^[21]将大规模知识库中已知实体的丰富同义名称作为训练数据，开发新型分层神经网络架构来预训练每种类型的实体匹配模型。Koutras等^[22]提出将数据库行、列和模式信息构建为图，再通过随机游走生成序列语料对模型进行训练，从而将图信息转换为可以揭示语义相似性的多维向量，并基于特征向量捕捉属性间的语义相似性来实现模式匹配。

虽然图结构能够表达关系型数据丰富的内在特征，但是图结构往往具有错误敏感性，真实关系型数据集中的脏数据容易误导基于图结构的数据特征学习，从而导致匹配效果不佳。因此，本文提出了基于模式匹配和预训练词向量的端到端的数据融合模型，在图的构建阶段只采用了数据库表结构信息，不考虑将数据实例纳入图中，以降低数据质量低下带来的负面影响。

2 基于图结构的数据融合模型

本文设计了一套端到端的数据融合模型，通过表间主外键和表字段关系将关系型数据库结构构建成图结构，基于图神经网络将任意两个图结构的节点进行匹配，实现两个数据源间的模式匹配，再将已匹配上的字段进行数据值标准化转换，最后将每阶段输出的结构转为可执行的SQL语

句，即实现了数据融合的核心步骤。本文提出的数据融合模型框架如图2所示，主要分为4步，每步形成流水线加工体系，各步的算法均可灵活替换，详细描述如下。

(1) 图的构建

给定一个关系数据库字典(dictionary, D)， D 中包含 n 张表(table, T)， $T=\{T_1, T_2, \cdots, T_n\}$ 为数据库 D 的所有表的集合，每张表包含 m 个字段属性(attribute, A)， $A=\{A_1, A_2, \cdots, A_m\}$ 为表 T 的属性个数，每张表的属性个数可能不相同，基于数据库中数据模型的表间主外键关系和表字段关系将数据库中表与字段关系构建成图 G (graph)，图中节点分为表和字段属性两类实体。如图2第一步所示，给定两个不同的数据集 D_1 和 D_2 ，可构建成 G_1 和 G_2 两个图。

(2) 模式对齐

数据库结构表级和属性级映射对齐，目的是将多个异构的数据源模式映射为标准模式，如图2第二步所示。将两个数据源中的人员基本信息表对齐，同时将两个表中包含的字段也对齐。通过将关系数据结构转为图输入DGMC模型，基于DGMC模型进行图模型训练，判断不同数据源表和字段属性实体之间的语义关联。DGMC模型本质上是通过学习图的向量表示，以及节点和关系的相似度，输出不同数据源表和字段属性的映射关系。

(3) 属性值对齐

数据实例标准化是将不同数据源同一属性的不同属性值做标准化转换，实现数据实例值与数据标准值映射对齐，如图2第三步所示，将学历值“本科”映射为标准值“大学本科”。考虑到预训练语言模型强大的语义表征能力和数据值短文本特性，本文采用预训练模型对需要进行标准化转换的单元格值进行标准化分类，实现

不同数据源属性值与数据标准值映射对齐。

(4) 生成SQL语句

将第二、三步的输出结果串联起来，生成数据库可执行的SQL语句，从而实现数据融合，如图2第四步所示。

2.1 图的构建

因缺乏统一的开发标准，不同业务系统的数据库模型各不相同，具体差异主要体现在事实表和维度表的设计、命名方式、数据字典的表现形式等方面。数据库字典通常以PDF、CHM、Excel等文件形式记录，预处理阶段需要将不同格式的数据库字典标准化为统一的图结构数据，通过图结构表示关系数据集中丰富的结构关系和语义信息。此步骤的输入数据为各业务系统的数据库字典，根据数据库字典中描述的表和字段的关系将其构建为网络图结构 G (graph)，数据库结构构建成图的实例如图3所示。其中表和字段均视为 G 的节点 V (vertex)，节点类型分为表和字段两种类型，基于数据模型主外键关系和表与字段间包含关系构成 G 的边 E (edge)，该步骤的输出为 $G(V, E)$ ，图 G 将作为模式对齐阶段的输入。

2.2 模式对齐

模式对齐是将两个系统的表、字段进行匹配对齐，本文将模式对齐问题转换为图对齐问题，即将两个图中的相关节点建立对齐关系。将需要进行对齐的图 G_1 和图 G_2 数据输入DGMC网络模型中，模型将输出两图节点的对齐结果。该模型的输入是待对齐的图 $G_1=(V_1, A_1, X_1, E_1)$ 、 $G_2=(V_2, A_2, X_2, E_2)$ ， V 表示节点集合， A 表示邻接矩阵， X 表示节点特征矩阵， E 表示

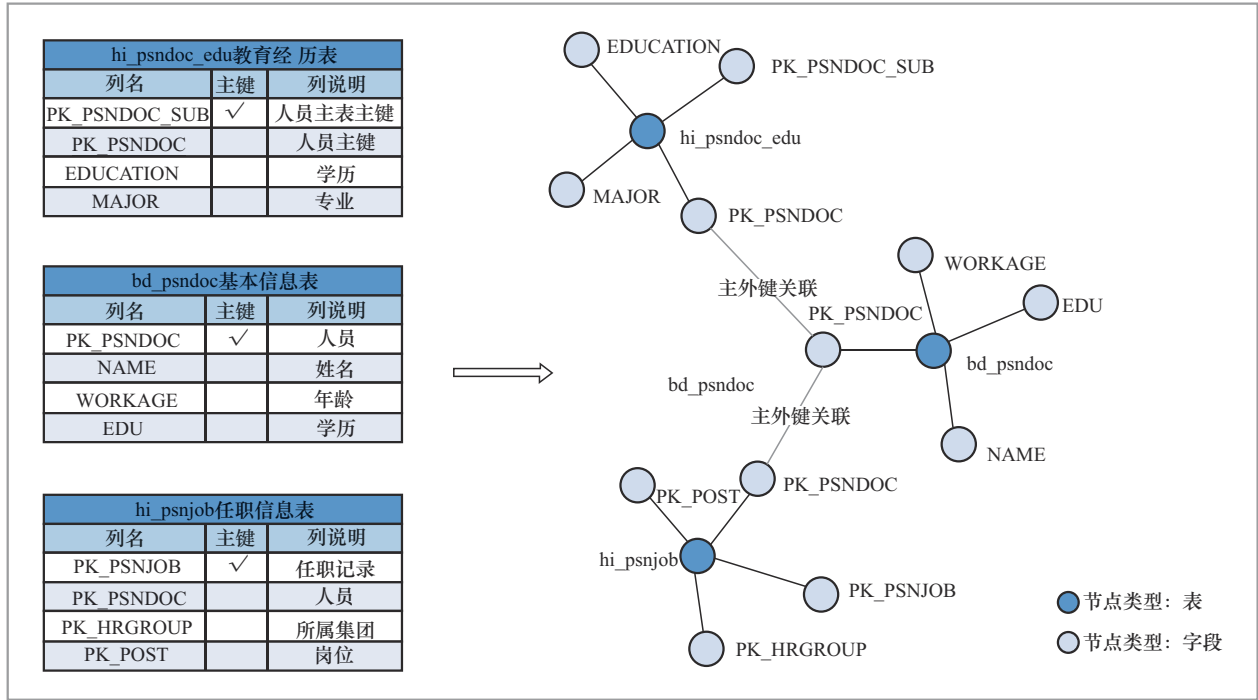


图3 数据库结构构建成图的实例

边的特征矩阵。其中，为充分利用图的特征信息，笔者采用表和字段中英文名称的预训练词向量 Glove^[23] 分别表示每个节点 V 的特征向量，使用机器翻译模型^[24] 将中文名称翻译成英文，并将中英文词向量进行拼接来表示节点的特征，记为 $X_{n \times m}$ ， n 为节点个数， m 为拼接后的向量维度。模型主要分为以下两个阶段。第一阶段主要基于节点相似度得到初始对应关系矩阵 $S^{(0)}$ 。首先利用共享权重的图神经网络 ψ_{θ_1} 提取图 G_1 和目标图 G_2 的深层节点特征 H_1 和 H_2 ：

$$H_1 = \psi_{\theta_1}(X_1, A_1) \in \mathbb{R}^{|V_1| \times \cdot} \quad (1)$$

$$H_2 = \psi_{\theta_1}(X_2, A_2) \in \mathbb{R}^{|V_2| \times \cdot} \quad (2)$$

其中， $\mathbb{R}$ 表示实数集合， $\cdot$ 表示每个节点的特征维度。

然后使用 Sinkhorn 算法^[25] 进行熵正则

化，得到初始的对应关系矩阵 $S^{(0)}$ ：

$$S^{(0)} = \text{Sinkhorn}(\hat{S}^{(0)}) \in [0, 1]^{|V_1| \times |V_2|} \quad (3)$$

$$\hat{S}^{(0)} = H_1 H_2^T \in \mathbb{R}^{|V_1| \times |V_2|} \quad (4)$$

第二阶段主要对对应关系矩阵 $S^{(l)}$ ， $l \in \{1, \dots, L\}$ 进行迭代优化，以达成局部最优的匹配结果。第一阶段得到的初始对应关系矩阵 $S^{(l)}$ 实际上是一个从源图的节点函数空间 $L(G_1)$ 到目标图节点函数空间 $L(G_2)$ 的一个映射，因此可得：

$$\vec{x}_2' = S^T \vec{x}_1, \vec{x}_1' = S \vec{x}_2 \quad (5)$$

然后以单位矩阵 $I_{|V_1|}$ 的形式构建源图的节点指示函数，利用图神经网络 ψ_{θ_2} 计算图特征，使用 ψ_{θ_2} 对 $S^{(l)}$ 对应关系矩阵进行优化更新，经过若干轮优化最终得到映射矩阵，输出实体对齐结果：

$$\boldsymbol{O}_1=\psi_{\theta_1}\left(\boldsymbol{I}_{\left|V_1\right|}, \boldsymbol{A}_1, \boldsymbol{E}_1\right), \boldsymbol{O}_2=\psi_{\theta_2}\left(\boldsymbol{S}_{(l)}^{\mathrm{T}} \boldsymbol{I}_{\left|V_1\right|}, \boldsymbol{A}_2, \boldsymbol{E}_2\right)$$

(6)

$$\boldsymbol{S}_{i, j}^{(l+1)}=\text { Sinkhorn }\left(\hat{\boldsymbol{S}}^{(l+1)}\right)_{i, j}$$

(7)

$$\hat{\boldsymbol{S}}_{i, j}^{(l+1)}=\hat{\boldsymbol{S}}_{i, j}^l+\psi_{\theta_3}\left(\vec{\boldsymbol{O}}_i^1-\vec{\boldsymbol{O}}_j^{(2)}\right)$$

(8)

图 4 所示为两个不同人力系统的数据库表和字段的对齐结果。模型输出的映射关系字典可能并不是完全准确的，会存在映射置信度低、映射缺失和映射失败的数据项。为了提高指标对齐的准确率，在实际应用场景中，可对输出的对齐结果进行人工校正。在人工复核的过程中，仔细核查置信度低、映射缺失和映射失败的数据项的映射结果，并根据相关的业务知识和经验对未对齐的数据项进行必要的修正和调整，通过复核再次输出数据项的对齐结果，以提高映射的准确性，确保对所有数据项都进行了适当的映射。

2.3 属性值对齐

该阶段主要对数据值进行标准化的映射匹配，其主要目的是将各系统字段值的不同维度值统一为相同值，为下游数据分析统计等提供标准化的数据。考虑

到实际场景中需要进行对齐的字段值通常是维度指标，维度值一般为较短的文本描述，因此，本文将数据实例中单元格对齐视为短文本分类任务，该任务基于预训练 Debertav3 模型来实现。Debertav3 模型是一种基于 BERT^[26] 结构进行跨语言任务处理的模型，通过在大规模语料上进行预训练和微调，能够有效地理解 and 处理多语言的文本数据，因此，本文不用标注任何数据对 Debertav3 模型进行微调训练。

本文在采用 Debertav3 模型完成分类任务时，将待对齐系统对应字段的枚举值文本作为模型的输入、数据标准字段的枚举值集合作为分类的标签（label），该模型能够将不同字段枚举值映射为对应数据标准的枚举值。经过实验测试，基于 Debertav3 模型实现属性值对齐，能够获得较高的准确率。图 5 所示为属性值与数据标准映射对齐，各企业基本信息表中的学历需要根据数据标准中的学历值进行转换，如企业人员信息表中的“专科”需要映射为数据标准中的“大学专科”，“本科”需要映射为“大学本科”，“小学、初中、高中”需要映射为“高中及以下”，同理，证件类型字段需要进行同样的转换。

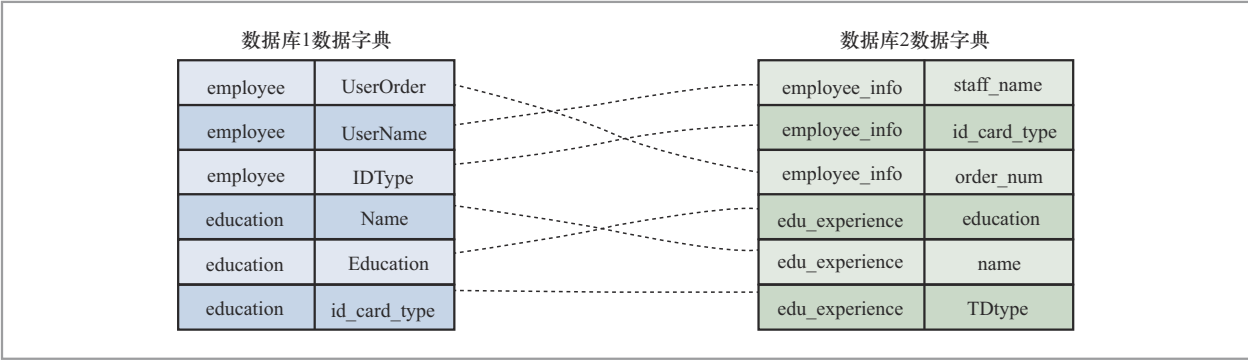


图 4 表和字段实体对齐实例

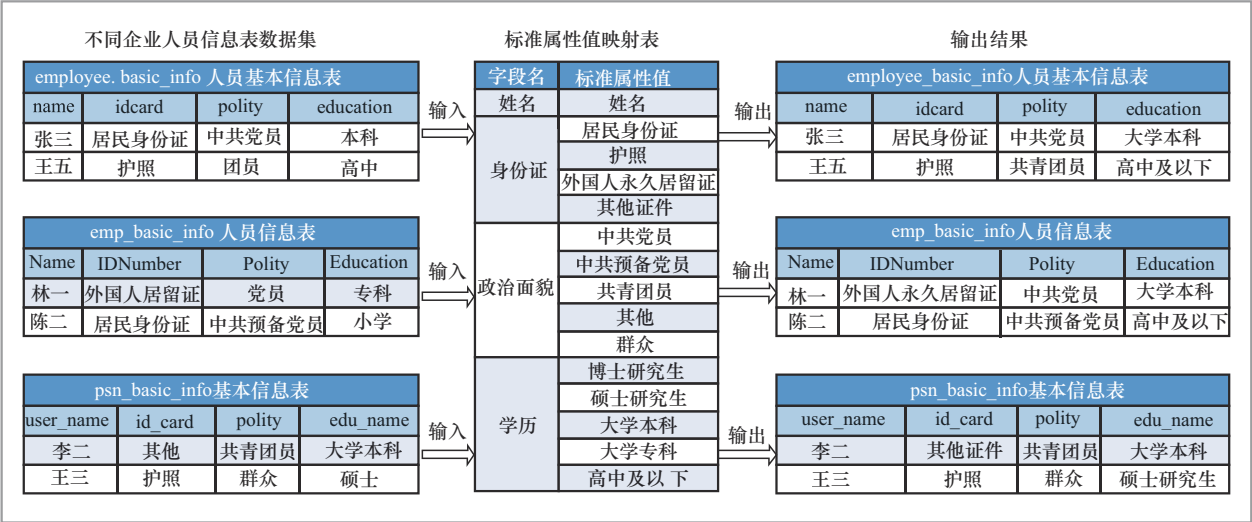


图5 属性值与数据标准映射对齐

2.4 生成 SQL 语句

根据第二步和第三步的结果，自动生成完整的数据融合 SQL 语句，SQL 语句的生成主要是经过以下两个步骤。

(1) 条件语句的生成

首先根据数据值的对齐结果，对每个对齐的字段生成“case when then”条件语句，此时若待对齐的字段或表名为空，则将“null”作为字段值。

(2) 完整语句的生成

通过获取待对齐表和字段的关系字典来获取表名，将“insert into”“from”添加到第一步的结果，最终输出数据库可执行的 SQL 语句。

下面用人力领域基本信息表实例来说明。经过模式对齐，可以将各个子企业系统中的基本信息表定位出来，同时通过属性值对齐找到待对齐基本信息表中各字段值与数据标准值映射关系，最终通过工程化规则生成 SQL 语句，数据融合 SQL 实例如图6所示。

根据每个需要对齐的系统表名生成相应的数据融合 SQL 语句，以便后续进行数据处理和分析，自动生成 SQL 语句减少了

```
insert into employee_basic_info
select
  hr.emp_basic_info.id, --序号
  hr.emp_basic_info.name, --姓名
  hr.emp_basic_info.age, --年龄
  case
    when hr.emp_basic_info.education == '硕士' then '硕士研究生'
    when hr.emp_basic_info.education == '高中' then '高中及以下'
  end as academic_degree, --学位
from hr.emp_basic_info

union all

select
  hr.psn_basic_info.id, --序号
  hr.psn_basic_info.name, --姓名
  hr.psn_basic_info.age, --年龄
  case
    when hr.psn_basic_info.education == '本科' then '大学本科'
    when hr.psn_basic_info.education == '初中' then '高中及以下'
  end as degree, --学位
from hr.psn_basic_info
;
```

图6 数据融合 SQL 实例

手动处理的工作量，并提高了数据的一致性和规范性。此外，生成的 SQL 语句也为后续实现数据值对齐所需的系统间数据整合提供了方便。

3 实验结果及分析

3.1 数据集

本文数据集来自企业数据融合项目的

实际场景，实际场景中集团型企业需要采集多个成员企业的人力系统、财务系统、党务系统等系统数据。为了建立各成员企业同业务领域不同数据库系统中表和字段的自动匹配关系，本次实验主要抽取人力业务领域的系统进行数据标注，共涵盖了6个不同数据库系统的人力系统数据。为了测试模型的泛化能力，也抽取了其他类型的系统数据进行标注，如企业的填报系统数据。数据预处理过程中包括数据字典清洗整理、表逻辑关系梳理、数据值标准规范映射。在模式对齐阶段笔者共对76张事实表、维度表及字段进行标注，标注过程需要对不同数据库表和字段建立匹配关系，匹配关系包括表和字段的关联关系。对于每一个字段，在另外一个数据库表中找到与之对应的字段，形成对齐的实体对，将表和字段之间的包含关系表示为图的边关系。经过标注后，该数据集包含1 944个实体和1 886条边。在属性值对齐阶段，笔者为868个数据项制定数据标准，其中需要进行属性值对齐映射的有60个数据项，以便模型对属性值进行标准化处理，数据集实例见表1，各数据库表、字段、数据值均需要与标准化标注进行映射。

3.2 评价指标

在模式对比阶段，本文采用3个评价指标对本文所提框架进行评测，分别是Hits@K(K≥1)、平均排名（mean rank，MR）和平均倒数排名（mean reciprocal rank，MRR），分别如式（9）、式（10）、式（11）所示。

$$\text{Hits}@K = \frac{1}{|S|} \sum_{i=1}^{|S|} I(\text{rank}_i \leq k) \tag{9}$$

$$\text{MR} = \frac{1}{|S|} \sum_{i=1}^{|S|} \text{rank}_i = \frac{1}{|S|} (\text{rank}_1 + \text{rank}_2 + \dots + \text{rank}_i) \tag{10}$$

$$\text{MRR} = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{\text{rank}_i} = \frac{1}{|S|} \left(\frac{1}{\text{rank}_1} + \frac{1}{\text{rank}_2} + \dots + \frac{1}{\text{rank}_i} \right) \tag{11}$$

其中，S是实体对集合，|S|是实体对集合个数，rank_i指第i个实体的预测正确的排名，I(·)是指示函数（indicator function），若rank_i≤k条件为真，则I(·)为1，否则为0。

Hits@K(K≥1)表示排名前K的实体对中正确匹配的实体占真值（真正匹配的实体对的数量）的比例，Hits@K的值越高，

表1 数据集实例

标准化标注			数据库1			数据库2		
数据项	字段名	标准值	表名	字段名	数据值	表名	字段名	数据值
工号	psn_no		emp_detail	emp_no		staff_base	base_code	
姓名	psn_name		emp_detail	name		staff_base	staff_name	
性别	gender	男、女	emp_detail	sex	M:男;W:女	staff_base	sex	1:男 2:女
年龄	age		emp_detail	age		staff_base	age	
婚姻状况	marital_status	未婚、已婚、 离异、其他	emp_detail	marital_status	未婚、已婚、 丧偶、离婚	staff_base	marriage	未婚、已婚、 丧偶、离婚

则实体解析的精度越高。本文使用 Hits@1、Hits@5 和 Hits@10 指标来评估模型的性能。MR 衡量了实体对齐结果中正确匹配的实体在排序列表中的平均位置，对于每一对实体对齐结果，记录其正确匹配的实体在排序列表中的位置，计算所有正确匹配的实体位置的平均值，MR 值越小，表示实体对齐的结果较好。MRR 计算所有正确匹配实体的倒数值值的平均值，即将每个正确匹配的实体位置的倒数值取平均。MR 和 MRR 都用于衡量实体对齐结果的准确性和排序性能，MRR 值越大，表示实体对齐的结果较好。

3.3 实验结果和分析

在模式对齐阶段分别对 Flex-matcher、BootEA 和 DGMC 模型进行了实验，并比较了 3 个模型在本文数据集上的实验效果，评测结果见表 2。由表 2 可以看出，DGMC 模型的 Hits@1、Hits@5 和 Hits@10 得分分别是 82.33、88.78 和 89.45，均高于其他两个对比模型。MR 和 MRR 得分分别为 1.37 和 0.91，而 Flex-matcher 和 BootEA 模型对应的 MR 和 MRR 得分分别为 13.46、0.16 和 8.67、0.24，可知 DGMC 模型的 MR 均小于其他两个对比模型，同时 MRR 值均大于其他两个对比模型。实验结果表明，DGMC 模型在 Hits@N、MR 和 MRR 指标上均取得最优的效果。图 7 所示为 DGMC 模型评测结果，由图 7 可知，训练过程中模型的迭代次数（epoch）在 140 左右时，模型的性能趋于平缓，可终止迭代。

基于实验结果，本文对 Flex-matcher、BootEA 两个对比模型的实验效果进行简要分析。Flex-matcher 模型是用于数据集成的工具，能够将多个数据源的

表 2 评测结果

模型	Hits@1	Hits@5	Hits@10	MR	MRR
Flex-matcher	14.67	31.46	42.63	13.46	0.16
BootEA	15.87	35.14	62.16	8.67	0.24
DGMC	82.33	88.78	89.45	1.37	0.91

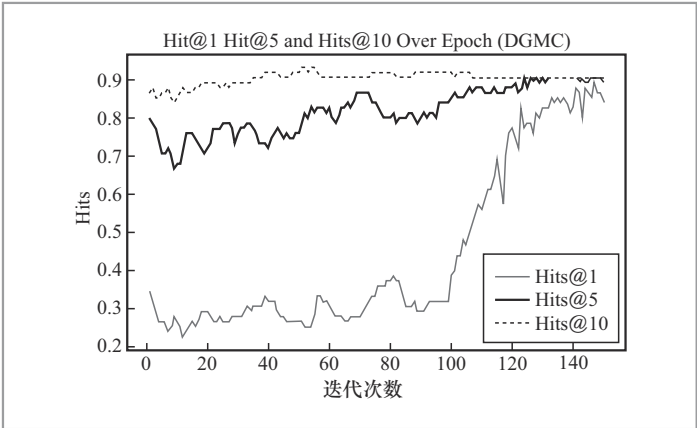


图 7 DGMC模型评测结果

表结构映射到统一的表结构中。然而，由于 Flex-matcher 模型采用传统的 KNN、N-Gram 等作为分类器，会存在数据不平衡、对异常值敏感等问题。此外，Flex-matcher 模型仅适用于同表的多个数据源间的映射匹配，泛化能力较差，且预测效果较差。BootEA 模型是基于嵌入的实体对齐方法，通过 TransE 来获取图结构属性，将不同系统的嵌入都统一到一个向量空间中，能在一定程度上捕捉到实体间的语义信息和图节点间相似性，但其建模能力较弱，并且基于 TransE 翻译假设的 BootEA 模型可能会将实体间的语义相似性视为结构相似性，导致模型给出不准确的对齐结果，预测效果也较差。

本文提出的数据融合框架集图构建、模式对齐和属性值对齐等为一体，在数据融合任务上具有一定的有效性，它能够根

据不同系统的数据结构和数据字典自动实现表和字段的对齐，进一步生成相应的SQL语句，从而实现数据的一致性和统一性，对于处理异构数据具有重要意义。

4 结束语

本文提出的基于图的异构数据融合模型能有效地解决不同系统之间的数据集成问题，该模型将数据融合拆分为图的构建、模式对齐、属性值对齐和生成SQL语句4个组件，每个组件均可用不同的模型进行替换，灵活性较强，后续可尝试用更多的算法替换不同的组件。该模型引入了图神经网络，通过关联数据库中各表的字段来构建图数据，输入图神经网络模型来实现不同系统的表和字段的自动匹配，能够将各业务系统中指代同一实体的记录连接起来。基于预训练语言模型将不同系统中的属性值按照数据标准进行映射分类，能够消除不同系统中同一实体属性值的差异，最后输出数据合并的SQL语句。实验结果表明，该模型在处理异构系统数据融合问题上具有较高的准确性，相较传统的手工融合，该模型能够在一定程度上提高数据融合的效率，减少手工处理的工作量，对企业数据融合具有重要的借鉴意义，并且不受领域限制，具备广泛的适用性。

本文仅通过数据库模式关系构建了网络图数据进行训练，但现实场景中数据库表间的逻辑关系不完整，需要人工梳理数据库表间的逻辑关系，该任务需要具备一定的技术基础，数据获取存在一定的难度。下一步可将每个列的属性值引入图中，以丰富图数据的语义关系，减少对数据库模式的依赖，通过充分利用已有数据值，进一步提高数据融合模型的性能。

参考文献：

[1] 吴信东, 应泽宇, 盛绍静, 等. 数据中台框架与实践[J]. 大数据, 2023, 9(6): 137-159.
WU X D, YING Z Y, SHENG S J, et al. Data central-platform: architecture and practice[J]. Big Data Research, 2023, 9(6): 137-159.

[2] 梅宏, 杜小勇, 金海, 等. 大数据技术前瞻[J]. 大数据, 2023, 9(1): 1-20.
MEI H, DU X Y, JIN H, et al. Big data technologies forward-looking[J]. Big Data Research, 2023, 9(1): 1-20.

[3] 闫佳和, 李红辉, 马英, 等. 多源异构数据融合关键技术与政务大数据治理体系[J]. 计算机科学, 2024, 51(2): 1-14.
YAN J H, LI H H, MA Y, et al. Multi-source heterogeneous data fusion technologies and government big data governance system[J]. Computer Science, 2024, 51(2): 1-14.

[4] 高云君, 葛丛丛, 郭宇翔, 等. 面向关系型数据与知识图谱的数据集成技术综述[J]. 软件学报, 2023, 34(5): 2365-2391.
GAO Y J, GE C C, GUO Y X, et al. Survey on data integration technologies for relational data and knowledge graph[J]. Journal of Software, 2023, 34(5): 2365-2391.

[5] CAPPUZZO R, PAPOTTI P, THIRUMURUGANATHAN S. Creating embeddings of heterogeneous relational datasets for data integration tasks[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2020: 1335-1349.

[6] FEY M, LENSSEN J E, MORRIS C, et al. Deep graph matching consensus[C]//Proceedings of ICLR 2020 Conference Blind Submission, Addis Ababa: Open-Review.net. 2020: 26-30.

[7] HE P C, GAO J F, CHEN W Z. DeBERTaV3: improving DeBERTa using

- ELECTRA-style pre-training with gradient-disentangled embedding sharing[EB]. arXiv preprint, 2021, arXiv: 2111.09543.
- [8] XIAO C, WANG W, LIN X M. Ed-Join: an efficient algorithm for similarity joins with edit distance constraints[J]. Proc VLDB Endow, 2008, 1: 933-944.
- [9] JACCARD P. The distribution of the flora in the alpine zone.1[J]. New Phytologist, 1912, 11(2): 37-50.
- [10] WU R Z, CHABA S Y, SAWLANI S, et al. ZeroER: entity resolution using zero labeled examples[C]//Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2020: 1149-1164.
- [11] COLLOBERT R, BENGIO S. SVM-Torch: support vector machines for large-scale regression problems[J]. Journal of Machine Learning Research, 2001, 1(2): 143-160.
- [12] CHEN C, BEHZAD G, ALON Y, et al. BigGorilla: an open-source ecosystem for data preparation and integration[J]. IEEE Data Engineering Bulletin, 2018, 41: 10-22.
- [13] BORDES A, USUNIER N, GARCIA-DURÁN A, et al. Translating embeddings for modeling multi-relational data [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. New York: ACM, 2013: 2787-2795.
- [14] WANG Z, ZHANG J W, FENG J L, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. New York: ACM, 2014: 1112-1119.
- [15] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2015, 29(1): 2181-2187.
- [16] JI G L, HE S Z, XU L H, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Stroudsburg: Association for Computational Linguistics, 2015: 687-696.
- [17] SUN Z Q, HU W, ZHANG Q H, et al. Bootstrapping entity alignment with knowledge graph embedding[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2018: 4396-4402.
- [18] WANG Z C, LYU Q S, LAN X H, et al. Cross-lingual knowledge graph alignment via graph convolutional networks [C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2018: 349-357.
- [19] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks[EB]. arXiv preprint, 2016, arXiv: 1609.02907.
- [20] WU Y T, LIU X, FENG Y S, et al. Relation-aware entity alignment for heterogeneous knowledge graphs[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 5278-5284.
- [21] ZHAO C, HE Y Y. Auto-EM: end-to-end fuzzy entity-matching using pre-trained deep models and transfer learning

ing[C]//Proceedings of the WWW19: The World Wide Web Conference. New York: ACM, 2019: 2413–2424.

[22] KOUTRAS C, FRAGKOULIS M, KATSIFODIMOS A, et al. REMA: graph embeddings-based relational schema matching[C]//Proceedings of the Workshops of the EDBT/ICDT 2020 Joint Conference. Copenhagen: CEUR Workshop Proceedings, 2020: 14–19.

[23] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: Association for Computational Linguistics, 2014: 1532–1543.

[24] JÖRG T, SANTHOSH T. OPUS-MT—Building open translation services for the world[C]//Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Lisboa: EAMT, 2020: 479–480.

[25] CUTURI M. Sinkhorn distances: light-speed computation of optimal transport [C]//Proceedings of the 26th International Conference on Neural Information Processing Systems – Volume 2. New York: ACM, 2013: 2292–2300.

[26] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019: 4171–4186.

作者简介



黄跃珍（1973– ），男，博士，广州数字科技集团有限公司党委书记、董事长，主要研究方向为多模态身份识别、公共数据运营、人工智能等。



杨芬（1991– ），女，广州广电运通信息科技有限公司数据分析师，主要研究方向为数据治理运营、自然语言处理、知识图谱等。



田丰（1975– ），男，广电运通集团股份有限公司研究总院院长，主要研究方向为数字底座技术、人工智能、大数据等。



张承业（1985-），男，博士，广州广电运通信息科技有限公司总经理助理，主要研究方向为数据底座技术、大数据平台技术、政企数字化转型研究等。



李雨婵（1998-），女，广州广电运通信息科技有限公司数据分析师，主要研究方向为数据治理、深度学习、可视化分析等。

收稿日期: 2024-04-23
通信作者: 张承业, zcyel@grgbanking.com
基金项目: 广州市黄埔区国际科技合作项目(No.2021GH12)
Foundation Item: International S&T cooperation program in Huangpu District of Guangzhou(No.2021GH12)