**文章编号**：0253-374X(2022)10-1392-13

**DOI**：10.11908/j.issn.0253-374x.22228

# 基于半监督学习的多源异构数据治理

饶卫雄[1]，高宏业[1]，林　程[1]，赵钦佩[1]，叶　丰[2]

（1. 同济大学 软件学院，上海 201804；2. 复杂系统仿真总体重点实验室，北京 100101）

**摘要**：为实现不同数据管理系统之间的互通，提出一种基于半监督学习算法的多源异构数据治理框架，并由此设计、实现和测试了一套非结构化数据与结构化数据的自动化对齐方法。利用命名实体识别(NER)技术，将非结构化数据转化为结构化数据，再分别利用基于字符串相似度的方法和基于监督学习的方法，对结构化数据进行模式匹配；通过半监督学习方法，在结构化数据与数据库记录实体之间进行实体匹配与融合；利用自然语言处理(NLP)技术及深度学习方法，对融合后的数据集进行缺失值填补。结果表明：在论文数据集和视频元数据集上进行对齐处理后，两者的F1值分别达到89.70%及96.50%；在不同属性上进行缺失值填补后，整体填补准确率达到78%以上，大大优于基线方法的准确率。

**关键词**：半监督学习；数据治理；多源异构数据；缺失值填补；命名实体识别(NER)

**中图分类号**：TP391　　**文献标志码**：A

## Multi-source Heterogeneous Data Governance Based on Semi-supervised Learning

*RAO Weixiong*[1], *GAO Hongye*[1], *LIN Cheng*[1], *ZHAO Qinpei*[1], *YE Feng*[2]

(1. School of Software Engineering, Tongji University, Shanghai 201804, China; 2. National Key Laboratory for Complex Systems Simulation, Beijing 100101, China)

**Abstract**: In order to realize the intercommunication between different data management systems, we proposed a framework of multi-source heterogeneous data governance based on semi-supervised learning. Then, we designed, implemented and tested an automatic alignment method of unstructured data and structured data. The named entity recognition (NER) technology was firstly employed in the framework to convert the unstructured data into the structured one, and the string-similarity-based method and supervised-learning-based method were respectively used for the schema matching of structured data. With the semi-supervised learning method, the structured data and its corresponding entity in database were matched and integrated. Finally, natural language processing (NLP) technology and deep learning methods were used to impute missing values in the integrated dataset. It is shown that the F1-scores for the alignment on the paper dataset and video metadata set are 89.70% and 96.50%, respectively; and that the accuracy of missing value imputation on different attributes is all above 78%, which is a great improvement compared with the baseline methods.

**Key words**: semi-supervised learning; data governance; multi-source heterogeneous data; missing data imputation; named entity recognition (NER)

近年来，随着物联网、云计算、移动互联网的迅猛发展，大数据已经成为蕴藏巨大价值的重要社会资源。高质量数据是人工智能模型的基础。任何算法的准确率，都取决于数据的完善程度、丰富程度以及结构化程度。因此，信息共享与数据互通已经成为影响人工智能产业发展的重要因素。例如，在一个大型组织中，各个部门的业务和功能条块分割，数据常常以不同的粒度和形式存储于不同的计算机系统，相互之间难以进行有效沟通，形成了所谓的“数据孤岛”。如何有效打破“数据孤岛”、提高数据质量，是数据治理的关键问题之一。

所谓“数据孤岛”是指多源异构数据。多源异构数据中的“多源”是指数据来自多个数据源，而且它们的数据存储平台和方式各不相同；多源异构数据中的“异构”是指描述同一个实体的数据类型复杂，

收稿日期：2022-05-10
基金项目：上海市科技重大专项(2021SHZDZX0100)；中央高校基本科研业务费专项资金
第一作者：饶卫雄(1974—)，男，教授，博士生导师，工学博士，主要研究方向为深度强化学习应用、时空数据科学、移动计算等。E-mail：wxrao@tongji.edu.cn
通信作者：赵钦佩(1982—)，女，副教授，硕士生导师，工学博士，主要研究方向为机器学习、数据挖掘、模式识别。E-mail：qinpeizhao@tongji.edu.cn

论文拓展介绍

数据结构不一致。整合多源异构数据时，往往通过数据集成平台对多个数据来源进行统一处理，在物理层面和逻辑层面上消除异构数据之间的差异，实现统一的表示、存储和管理，将多源异构数据集成为相互理解、相互关联的有机整体，最终提升系统的数据处理效率。这一步又被称为多源异构数据治理[1]。目前，普遍采用如图1所示的方案[1-2]实现多源异构数据治理。

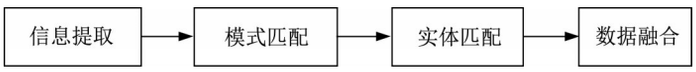


**图1　多源异构数据治理的普遍方案**

**Fig.1　Common framework of multi-source heterogeneous data governance**

该方案的主要支撑技术包括信息提取、模式匹配（又名模式对齐）、实体匹配（又名实体对齐）和数据融合。其中，信息提取及模式匹配的目的是处理数据本体的异构性和数据源的异构性[3]。实体匹配的目的是利用实体的属性信息构建对齐关系，同时对实体包含的信息进行聚集和融合。冲突解决是数据融合的必要过程。

在大规模数据集成处理项目中，通常采用的传统方法是利用人力来执行分类、集成、链接和聚合等操作。这种方法的缺点是代价过高。随着系统的拓展，基于人力的大数据集成将难以为继。在人工处理的基础上，工业界引入了主数据管理（master data management，MDM）系统，如 Informatica 和 IBM 提供的 MDM 系统。然而，这类系统的用户参与度高并依赖于人工规则，因此缺乏拓展潜力。机器学习（machine learning，ML）是一种具有良好可拓展性的方法。简单说，就是可以利用机器学习实现各个步骤[3]的自动化。

在信息提取方面，基于机器学习的信息提取主要用到3种方法，包括基于分类的提取、基于顺序标记的提取以及基于规则的提取[4]。基于分类的提取是将信息的提取转化为分类问题，即：通过检测特殊类型信息的边界以提取中间信息。基于这一思想的算法最初由 Finn 等[5]提出，使用少量已标记边界的数据分别训练开始边界分类器和结束边界分类器。该方法的缺陷也很明显，一是训练集中的负例数据量远大于正例数据量，二是在特定领域的信息提取上边界匹配不够准确。基于顺序标记的提取的目标是建立一个模型以供人工智能学习，通过标记观测序列进行预测。主流方法包括隐马尔科夫模型[6]、最大熵马尔科夫模型[7]、条件随机场[8]等。对于基于规则的提取，主要使用几个通用的规则从文本中提取信息。基于该方法的系统，如 LP[9]、Koko[10]等系统，依赖规则进行信息提取，但该方法往往受到数据获取和人工参与的限制。

在模式匹配方面，当前的研究主要分为基于语义相似度的模式匹配和基于机器学习的模式匹配。基于语义相似度的模式匹配着眼于2个属性值之间的句法相似度或语义相似度，通过将值转化为词向量的方式判断两者之间的相似度。该方法能够处理的数据类型较为单一，而基于机器学习的模式匹配则可以很好地解决这一问题。

实体匹配指在无相同主键的情况下将2个结构化数据表进行匹配的技术。目前主要的方法为基于属性值相似度的监督学习的匹配方法。Bilenko[11]提出了一种基于主动学习的实体匹配方法，通过少量标记数据训练人工智能以识别记录间相似度。Kopcke 等[12]指出：传统的机器学习模型，如随机森林模型等，显著改善了实体匹配的准确度；深度学习模型则通过词嵌入来比较长文本值，因此在文本及脏数据的匹配方面具有优势[13-14]。然而，最近的一项研究表明，在匹配一对数据集时，至少需要150万个标签才能够达到99%的精度与召回值[15]。不难发现，当缺乏足够的标记数据作为训练数据时，基于传统机器学习模型和深度学习模型的方法就会受到很大限制。针对这一问题，Blum 等[16]提出，可以利用半监督学习的方法来获得更多标签数据。

数据融合主要包括属性对齐及冲突解决两部分。属性对齐的目的是将对齐后的实体中能够映射为同一个属性的不同表述进行融合，从而获得更精确、更完善的信息。Cheatham 等[17]提出以属性描述文本为基础进行属性对齐。然而，所利用的知识库很可能并不含有所需的描述信息。基于数据驱动的对齐方法在一定程度上可以规避传统方法产生的问题。Yu 等[18]提出，可以通过属性函数的相似性计算出属性对齐的结果。冲突解决的目的是处理同一实体的同一属性出现语义相同但表述不一或语义冲突的情况，并对融合后的实体进行推演，利用已有的属性来推断另外的缺失属性值。主要方法包括基于统计的填补和基于学习的填补。刘莎等[19]提出了一种

基于灰色关联度的缺失值填补框架,其分类精度优于传统机器学习方法;Koren等[20]提出了将缺失值填补视为矩阵分解问题的填补思路。填补结果在稳定性、速度和准确性方面均有所提升,但是在处理包含自由文本的数据列时容易丢失一定量的有价值信息。

目前,基于机器学习的数据治理方法仍然需要大量人工标签数据。此外,数据融合方面的相关工作较少,尚未有一套基于中文的多源信息提取并结合了实体融合与推演的完整框架。

基于以上考虑,将信息提取、模式匹配、实体匹配以及数据融合各步骤相连接,提出了一个基于半监督学习的多源异构数据治理架构。首先对架构的基本思想以及4个步骤所用方法进行介绍,然后利用5个真实世界的数据集对该架构的各步骤进行验证与评估。

## 1 基于半监督学习的多源异构数据治理架构

基于半监督学习的多源异构数据治理架构如图2所示。

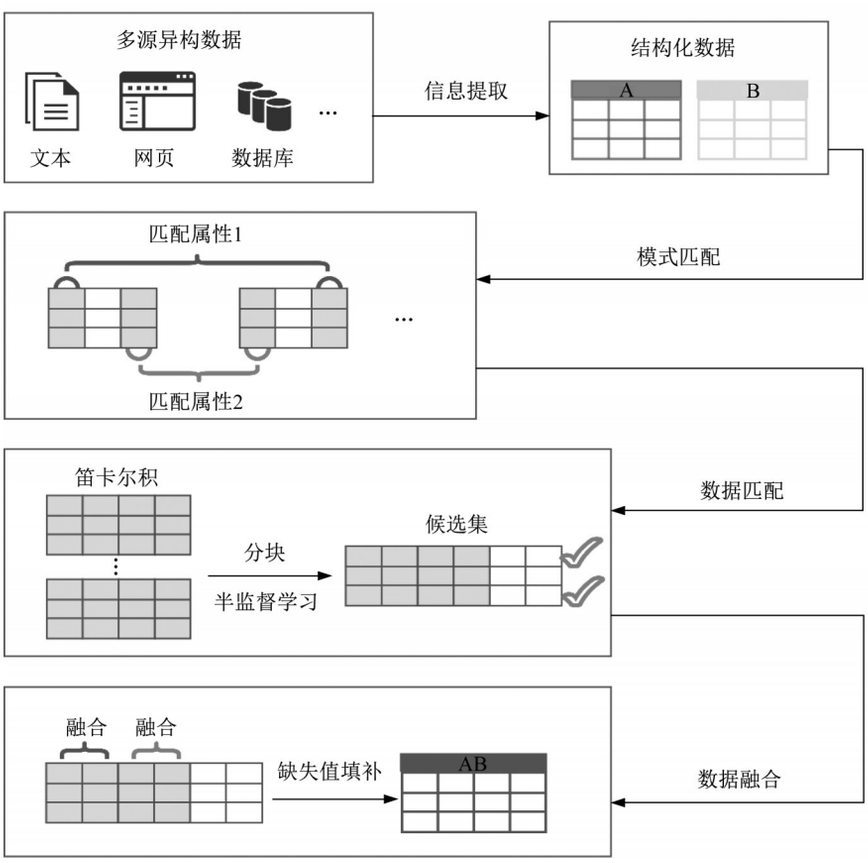


**图2 基于半监督学习的多源异构数据治理架构**

**Fig.2 Framework of multi-source heterogeneous data governance based on semi-supervised learning**

以图2展现的数据为例,描述同一实体的数据来自于不同的数据源,如非结构化的自然语言文本数据、半结构化的网页数据以及结构化数据库表数据等。首先,采用信息提取技术,将多源数据转化为统一的结构化数据库表形式;然后,通过模式匹配,找到结构化的数据库表之间匹配的属性对(匹配的列),再利用基于半监督学习的数据匹配技术,找到结构化数据库表之间匹配的记录(指向同一实体的行);最后,通过数据融合,将匹配的列属性值融合在一起,再进行一定程度的实体推演,完成缺失值填补。

### 1.1 信息提取

信息提取的目的是从文本数据或是网页数据中提取人们感兴趣的信息,目前应用较为广泛的信息提取技术是基于自然语言处理(natural language processing, NLP)的提取技术。鉴于信息提取对象的特殊性,在非结构化数据信息提取时,提取的准确性取决于自然语言处理的准确性。本研究中采用的

信息提取算法为：通过自然语言处理中的词性识别以及命名实体识别（NER）判别实体，再结合正则表达式和一定的规则提取特殊信息。基于自然语言处理的信息提取算法，虽然减少了人工制定提取规则的工作量，但是也存在缺陷。首先，基于自然语言处理的命名实体识别技术只可提取人名、地名、机构名、时间、日期、货币及百分比信息；其次，可提取的信息范围以及准确性有限，需要额外制定一些规则以去除错误的匹配信息；再次，识别范围有限，需用正则匹配提取一些无法通过自然语言处理来识别的复杂结构数据。值得一提的是，正则表达式增加了基于命名实体识别的信息提取算法的灵活性。基于命名实体识别的信息提取算法伪代码如图3所示。

```
Input: 非结构化数据UD,正则表达式re,规则R
Output: 数据表T;
  for data in UD:
    row=[]
    row.append(正则表达式re匹配出来的信息)
    NLPData=对输入的数据进行NLP处理，识别词性及命名实体
    如果data的NER tag为七类命名实体之一且符合规则R:
    row.append(data)
    T.append(row)
Return T
```

**图3　基于命名实体识别的信息提取算法（算法1）**

**Fig.3　Information extraction algorithm based on named entity recognition (algorithm 1)**

### 1.2　模式匹配

模式匹配的目的是从输入的多个属性互有重叠的表中找到相同的属性，其基本思路是：寻找不同列之间属性值的关联性，根据关联性判断两列属性是否相同。显然，模式匹配的关键在于求出匹配的属性名以及属性值之间的关联程度。本研究中采取了2种不同的方法完成对不同属性值类型的匹配，分别是基于机器学习方法的模式匹配和基于属性值相似度的模式匹配。

#### 1.2.1　基于机器学习方法的模式匹配

基于机器学习方法的模式匹配采用了多种机器学习方法，如使用数据表中的属性名和属性值训练分类器。本研究中使用FlexMatcher包中的KNN Classifier、*n*-Gram Classifier、CharDist Classifier及Flex Matcher进行基于属性值的训练，以获取2个待匹配数据表（表A和表B）中匹配的属性对。

KNN Classifier中将$k$定义为3，一个属性值点被归类为距离该点最近的3个邻近样本数据点中使用最频繁的一类。同时使用莱文斯坦距离作为距离度量。莱文斯坦距离是编辑距离的一种，指2个字符串之间由一个转换成另一个所需的最少编辑操作次数。

*n*-Gram Classifier考虑了所提取属性值的特征，使用单词或字符的*n*-grams作为数据处理的单元，其中$n$为自定义参数。通过提取转化出的*n*-grams的计数特征或哈希特征训练逻辑回归分类器。

CharDist Classifier从字符串中的特殊字符着手，提取数据中字符类型的计数，并以其为特征，共利用7种计数特征训练逻辑回归分类器。

Flex Matcher综合了上述分类器的预测结果，并对结果进行加权计算以获得最佳效果。

#### 1.2.2　基于属性值相似度的模式匹配

基于属性值相似度的模式匹配的基本思想是：将字符串间的相似度作为关联度的评判标准。具体操作方法为：通过判断不同属性值之间字符串的距离求出不同属性值之间的相似度。字符串的相似度有多种衡量方式，如杰卡尔德相似度、编辑距离等。本研究中主要使用杰卡尔德相似度度量字符串的相似性。

杰卡尔德相似度的定义为：2个集合$S_1$与$S_2$交集的元素个数除以$S_1$与$S_2$并集的元素个数。杰卡尔德相似度计算式如下所示：

$$J(S_1,S_2)=\frac{|S_1\cap S_2|}{|S_1\cup S_2|}$$

基于属性值相似度的模式匹配算法的伪代码如图4所示。该算法的基本思想是：选取A表一列属性attr1中的一个属性值value，计算其与B表某一列属性attr2中的所有属性值的杰卡尔德相似度，取其中最大值记为max；然后，对attr1中所有属性值执行相同操作，得到一个长度为A表行数的最大值数组；最后，取该数组平均值。该平均值就是A表中的属性attr1与B表中的属性attr2的相似度。对B表中的所有属性执行上述操作后，获得一个长度为B表列数的相似度数组，取其最大值对应的属性。若最大值不为零，则认为最大值对应的属性与attr1匹配；若所有属性的相似度都为零，则认为attr1属性没有匹配属性。

### 1.3　数据匹配

数据匹配的目的是找到2个表之间指向现实中同一实体的元组对。数据匹配的难点在于如何获取正确匹配的数据元组。本研究中采用半监督学习方法来筛

```
Input: 数据表A，数据表B
Output: 匹配属性对Matching;
  for attr1 in A:
    similarity=[]
      for attr2 in B:
        for value in attr1:
          计算value与attr2所有属性值的Jaccard最大值max
        similarity.append(所有max值的均值)
      Matching.append({attr1, similarity最大值对应的属性})
Return Matching
```

**图4 基于属性值相似度的模式匹配算法(算法2)**

**Fig.4 Pattern matching algorithm based on attribute value similarity (algorithm 2)**

选出最佳的匹配数据。数据匹配的流程如图5所示。

在获取了2个数据表中匹配的属性对后,需要找到2个数据表中指向同一实体的记录(匹配的行)。对数据表进行笛卡尔乘积,找到2个数据表中所有指向同一实体的记录(将2个表中被笛卡尔积连接在一起的一对记录称为元组对,将指向同一实体的元组对称为匹配元组),以免遗漏任何匹配元组。以A表、B表匹配为例。首先分别进行随机采样,生成A′及B′2个子表,然后对A′及B′2个子表生成笛卡尔积。在数据量十分庞大时,如果直接对2个数据表进行笛卡尔乘积,数据匹配的代价就会过高。因此,先选择一对匹配的属性对,然后在候选匹配实体表中移除一些明显不匹配的元组,这一步骤被称为分块。

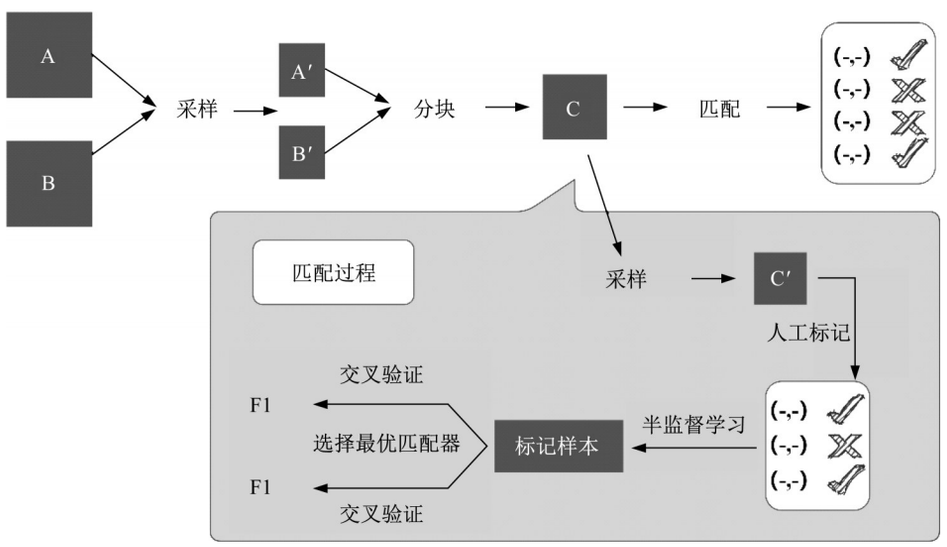


**图5 数据匹配流程**

**Fig.5 Flowchart of data matching**

此时需要找出C表中匹配的元组,这里采用半监督学习方法。本研究中使用的半监督学习方法是基于差异的半监督学习方法。因为有多种方式可表示机器学习中的数据特征,所以在协同训练方法的基础上,Zhou等[21]又提出了Tri-Training算法。Tri-Training算法的特点是:在原始数据集上随机抽取不同的训练集进行训练,以保证分类器的差异性。与协同训练方法不同的是,Tri-Training算法采用了3个分类器,即$R_i$、$R_j$、$R_k$。因此,可信标记数据就可由简单投票法则确定。具体做法是:如果分类器$R_i$和分类器$R_j$对于C表中的未标记记录$x$的标记是相同的,就把$x$及其标记$y$加入到$R_k(k\neq i,j)$的标记训练数据集中。

根据从模式匹配中获取的匹配属性对,为属性值生成一系列的特征,包括匹配属性对的属性值之间的余弦距离、编辑距离、杰卡尔德距离等衡量属性值相似度的特征。首先,从C表中取出一部分数据作为样例数据,再从样例数据中取出少部分数据,以人工方式标注它们是否匹配。为了最小化人工参与的程度,采用Tri-Training算法获得整个样例数据集合中的标签,并应用交叉验证,通过F1值选择最优的匹配方式。

### 1.4 数据融合

数据融合的目的是:将匹配属性对中的属性值进行融合。数据匹配工作完成后,多源数据集已横向合并为一个数据集,但公共属性集中的对应属性列还未融合。对于同一实体属性的不同属性值,应根据其数据特点进行属性合并。对于数值类属性,

可以根据数据的特点，采取平均值、中位数或其他一些统计值作为融合后的属性值；对于字符类属性，可以保留较长、较多的属性值，也可以保留其并集；对于一些应以某个数据集为准的属性，直接保留该数据集的属性值。

属性融合完成后，某表独有实体的非公共属性上产生新的缺失值。实际上，这部分缺失值的信息很可能仍然隐藏在该实体的公共属性中。因此，可以通过自然语言处理与数据填补相结合的方法进行一定程度上的实体推演，对实体融合后的缺失值进行填补。

对于数值型数据，通常可以将其看作一个回归或矩阵补全问题；对于字符型数据，在进行一些自然语言处理后，将其转化为分类问题。本研究中提出的字符型缺失数据的填补框架如图 6 所示。

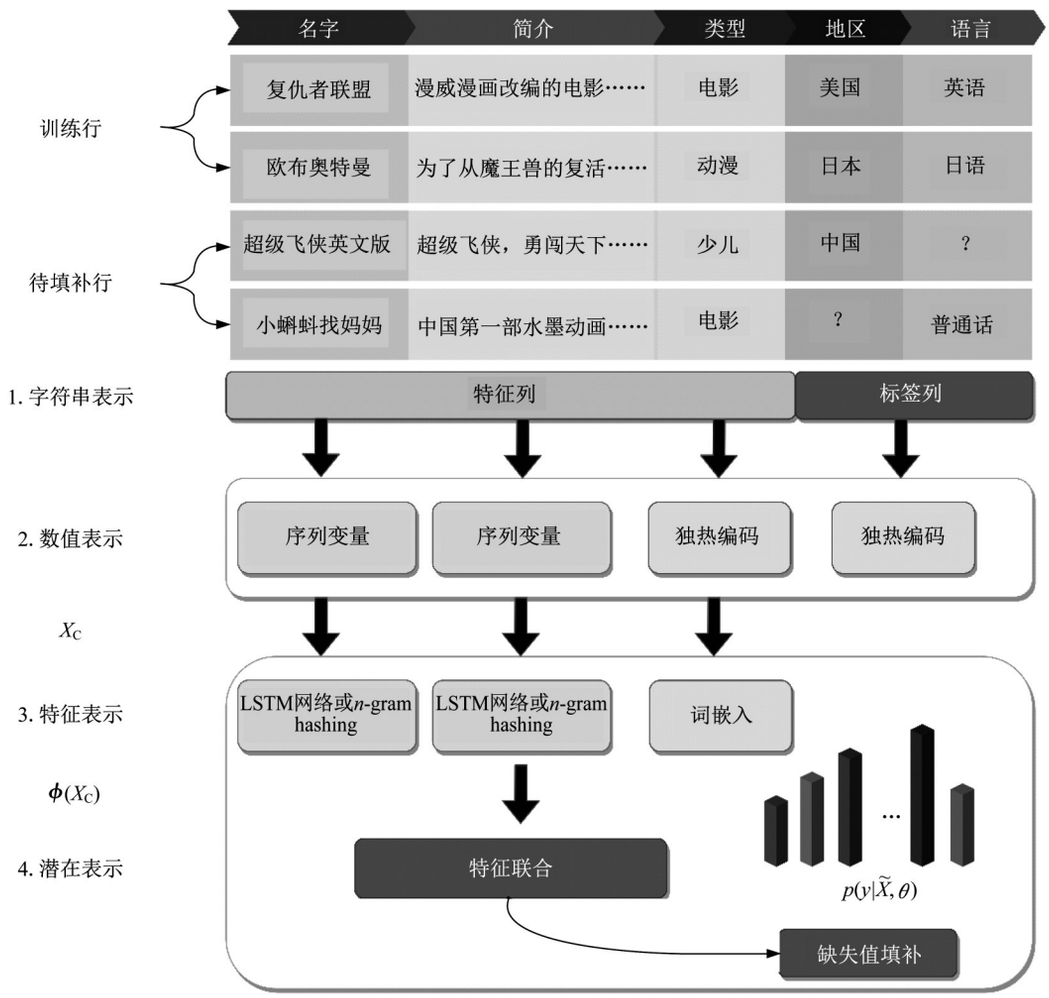


**图 6　字符型缺失数据填补框架**

**Fig.6　Framework of missing data imputation for character type**

对于实体融合后数据集中的缺失值，很可能在其他属性中被找到。例如，对于图 6 中待填充的元组“小蝌蚪找妈妈”，在原来的数据集中，“地区”属性值缺失，但实际上简介中包含了“中国”这个地区信息，因此预测该元组的“地区”可能为“中国”。简而言之，该框架先预测所有属性可能取到的值，然后取可能性最高的值作为该值的插补。具体流程分为以下 4 个阶段：

(1)字符串表示阶段。将数据列分为输入列(作为特征)和要插补的列(作为目标)，数据仍然是以文字形式表示。具有观测值的所有行均被视为训练数据(以及验证数据或测试数据)，对缺失值的行执行插补操作。

(2)数值表示阶段。根据列数据类型的不同，通过独热编码，将分类变量转化为数值。独热编码又称“一位有效编码”，具体指：用状态数相等的位数对每个状态进行编码，并且只有一位可以取到有效值。通过简单的词袋模型，将序列变量(如自由文本字段)转化为数值。本阶段生成的数值记为 $X_C$。

(3)特征表示阶段。机器学习模型的预测质量

主要取决于所使用的特征表示,本阶段以大量的嵌入工作为基础,将分类变量和序列变量的数值表示转化为可学习的特征表示。对于分类变量,通过词嵌入将独热编码转化为词向量;对于序列变量,采取LSTM网络及*n*-gram hashing这2种方法进行特征提取。LSTM网络又称长短期记忆网络,具体指:通过对上一阶段的词袋表征进行特征压缩,获得一个更有效的词向量表示。*n*-gram hashing指:首先利用*n*-gram模型对上一阶段的词袋表征进行维度扩充,然后利用哈希降维得到特征向量。本阶段的目的在于:提升$X_c$的质量,并将最终生成的特征向量记为$\boldsymbol{\phi}(X_c)$。

(4)潜在表示阶段。本阶段中将所有特征列的特征向量$\boldsymbol{\phi}(X_c)$串联为一个最终特征向量,再通过Softmax进行多分类,以便完成目标列缺失值的填补。

## 2 实验与评估

实验的目的是实现多源异构数据治理框架。如图2所示,分别对信息提取、模式匹配、数据匹配、数据融合进行评估,以验证该框架的有效性与准确性。实验共采用5个数据集。

(1)信息检索(information retrieval, IR)数据集。该数据集从国内某大学知识库网站中获取,记录了某高校所有的7 264名学者的个人信息。其中,“edu”为学者的学历,“title”为学者的职称,“intro”为学者的个人简介,“field”为学者的研究方向,“paperQuan”为学者的论文数量,“publication”为学者的论文,“coauthor”为学者论文的合著者。以上个人信息共同构成一个完整的结构化数据表。

(2)数据库系统和逻辑编程(database systems and logic programming, DBLP)数据集。该数据集从DBLP网站中获取。DBLP数据集是计算机领域内的一个英文文献的集成数据库系统,以作者为核心,按时间顺序列出了作者所有的科研成果,包括国际期刊和会议论文等所有公开发表的论文。该数据集的属性列包括部门、学历、职称、邮箱、dblp_papers及合著者。

(3)爱奇艺优酷腾讯元数据集。视频数据源来自爱奇艺、优酷及腾讯三方,涉及电影、电视剧、综艺、纪录片等视频类型。该数据集提供总计215 169条描述这些视频的元数据,包含名字、演员、导演、类型、发布年份等22个属性。

(4)智能电子节目指南(electronic program guide, EPG)补充元数据集。该数据集提供总计9 379条电影、动画等来自智能EPG的视频元数据,包含演员、导演、语言、地区、故事、标签、年份等25个属性。

(5)同洲媒资库元数据集。该数据集提供总计500 013条元数据,覆盖电影、电视剧、新闻、音乐MV、教育课程等视频类型,包含名字、演员、放映时长、地区、简介、年份等47个属性。

分别综合2个论文数据集及3个视频元数据集的特点,从中提取出一致的、更完善的信息,建立了2个更具凝聚力的数据集,再通过缺失值填补,提升融合数据集质量。然后,在多源异构数据治理的流程中,依序进行不同方法的对比。

### 2.1 信息提取实验

该实验中采用信息提取技术获取DBLP数据集的论文数据。其余数据集均为结构化数据表,因此无需进行信息提取。

DBLP数据集的获取过程如下:根据第1.1节中算法1,结合自然语言处理中的命名实体识别技术,从某高校机构知识库网站提取了学者的姓名、部门、职称和学历信息,然后利用正则表达式提取学者的邮箱信息。由于信息提取算法的效果格外依赖于命名实体识别的准确性,因此比较了5种中文自然语言处理命名实体识别在学者信息识别上的准确性。

(1)Google自然语言处理命名实体识别接口。一个由Google提供的支持多语言命名实体识别的应用程序编程接口。

(2)THULAC中文词法分析工具。一套由清华大学研制并推出的中文词法分析工具包,具有中文分词和词性标注等功能。

(3)DeepNLP工具。通过将Tensorflow深度学习平台上的模块与最新算法相结合,提供自然语言处理基础模块,并支持其他更加复杂的任务拓展。

(4)jieba分词工具。一个Python中文分词组件,支持中文文本的分词、词性标注、关键词抽取等操作;它的命名实体识别功能能够准确识别出人名,但不能准确识别出人名外的其他实体。

(5)HanLP工具。一个由一系列模型与算法组成的自然语言处理工具包,提供的命名实体识别技术支持人名、地名及机构名的准确识别,能够较好地识别出人名、职称、学历等信息。

综合上述工具包的实验效果,采用HanLP工具和jieba分词工具共同识别学者姓名,然后根据取较

长字符串的原则，融合两者的姓名识别结果。在使用HanLP工具识别学者的职称、学历及机构信息时，由于机构名的构成比较复杂，因此该工具只能完整识别出部分机构名，存在识别出不完整机构名的问题。该信息提取算法的信息提取准确性如表1所示。

**表1　信息提取算法准确性**

**Tab.1　Accuracy of information extraction algorithm**

| 各信息提取的准确性/% | | | |
|---|---|---|---|
| 姓名 | 职称 | 学历 | 部门 |
| 67.63 | 34.60 | 87.50 | 80.46 |

随后，以获取到的学者姓名拼音为关键词，结合DBLP网站提供的英文文献数据集中的论文数据与合著者数据，最终获得了实验中用到的DBLP数据集。

### 2.2　模式匹配实验

对上述5个数据集进行模式匹配，分别进行了2组实验。第1组实验针对论文数据集，将IR数据集中的属性匹配DBLP数据集中的属性；第2组实验针对视频元数据集，利用智能EPG补充元数据集中的属性来匹配爱奇艺优酷腾讯元数据集和同洲媒资库元数据集中的属性。本实验采用了第1.2节中介绍的模式匹配方法。

#### 2.2.1　论文数据集模式匹配

首先，在IR数据集及DBLP数据集上执行基于机器学习的模式匹配，各分类器匹配结果如表2所示。表2中，√表示正确匹配，×表示错误匹配。

**表2　DBLP数据集与IR数据集模式匹配结果**

**Tab.2　Pattern matching results between DBLP dataset and IR dataset**

| DBLP数据集属性 | 各分类器匹配结果 | | | |
|---|---|---|---|---|
| | KNN Classifier | CharDist Classifier | 3-Gram Classifier | Flex Matcher |
| 部门 | √ | √ | √ | √ |
| 邮箱 | × | × | × | × |
| 姓名 | √ | × | × | √ |
| 学历 | × | × | √ | √ |
| 职称 | √ | × | × | √ |
| 合著者 | × | √ | × | × |
| dblp_papers | × | × | × | × |

可以看出，3-Gram Classifier的模式匹配效果明显不如其他3种分类器。用KNN Classifier和CharDist Classifier进行模式匹配后，正确匹配不大于3个，而使用综合所有分类器的Flex Matcher时，正确匹配有4个。这是由于CharDist Classifier和KNN Classifier提取出的特征不能清晰地反映字符串属性值的特性。

采用绝对多数投票法时，由于Flex Matcher分类器属于集成分类器，其权重较大，因此本实验中将其权重设为2；若2个结果的票数相当，则取Flex Matcher的结果作为匹配结果。在对$n$-Gram Classifier进行绝对多数投票时，其内部的多个分类器先进行一次投票，投票结果作为$n$-Gram Classifier的结果。最终获得的匹配结果如表3所示。

**表3　DBLP数据集与IR数据集绝对多数投票结果**

**Tab.3　Majority voting results for DBLP dataset and IR dataset**

| DBLP数据集属性 | 模式匹配结果(IR) | 正确匹配属性(IR) | 投票 | 是否匹配 |
|---|---|---|---|---|
| 部门 | depart | depart | 5 | √ |
| 邮箱 | ins | email | 5 | × |
| 姓名 | name | name | 4 | √ |
| 学历 | edu | edu | 2 | √ |
| 职称 | title | title | 2 | √ |
| 合著者 | intro | coauthor | 3 | × |
| dblp_papers | intro | publication | 9 | × |

采用基于机器学习的模式匹配方法时，在实验数据集上的匹配结果，不如采用基于相似度的匹配方法的结果好。原因可能是：实验数据集中的数据基本都是字符串型数据，而计算字符串的相似度能够直接反映出2个属性值的相似度，因此基于相似度的模式匹配得到了更好的结果。

#### 2.2.2　视频元数据集模式匹配

实验中采用4种分类器对视频元数据集执行模式匹配，结果如表4所示。

**表4　视频元数据集与智能EPG数据集模式匹配结果**

**Tab.4　Pattern matching results between video metadata sets and smart EPG datasets**

| 智能EPG数据集属性 | 各分类器匹配结果 | | | |
|---|---|---|---|---|
| | KNN Classifier | CharDist Classifier | 3-Gram Classifier | Flex Matcher |
| 名字 | √ | × | √ | √ |
| 演员 | × | × | × | × |
| 导演 | × | × | × | × |
| 年份 | √ | √ | × | √ |
| 地区 | √ | × | √ | √ |
| 语言 | √ | √ | √ | √ |
| 标签 | √ | √ | × | √ |

可以发现，CharDist Classifier和3-Gram Classifier的匹配结果较差。CharDist Classifier仅使用一些特殊字符的计数特征进行训练，因此很难处理名字、地区等不含或者很少含有特殊字符的属性。与上述2种分类器相比，KNN Classifier及Flex

Matcher的效果稍好,但仍存在一些匹配的误差。

总体来说,基于机器学习的模式匹配,在既包含字符属性又包含数值属性的数据集上取得了良好的实验结果,其中Flex Matcher得到了最优匹配结果。因此,在对数据表进行模式匹配时,可以根据数据表中数据值的格式,有选择性地灵活使用上述方法,以获得较好的匹配效果。

### 2.3 数据匹配实验

在获取了匹配的属性对后,或者说,在完成了纵向对齐后,开始数据匹配实验。实验中将对2个表指向同一实体的记录进行横向对齐。本研究中采用Magellan包提供的分块方法执行对齐,并将结果进行合并,以保证不会错过正确匹配的元组。第1种分块方法是Attribute Equivalence Blocking,即获取所有在指定属性对上的值完全相等的元组;第2种分块方法是Overlap Blocking,即获取所有在指定属性对上的值有一定程度相同的元组;第3种分块方法是Rule Based Blocking,即通过人为制定相似度规则对元组对进行过滤。将这3种方法获取到的元组合并之后,即可获得一个排除掉明显不匹配元组的笛卡尔积数据表C。

#### 2.3.1 论文数据集数据匹配

首先,在DBLP数据集和IR数据集中选取合适的属性对,并对其执行分块操作。排除含有较多空值的(邮箱,email)/(职称,title)/(学历,edu)属性对,同时排除字符串长度过长的(dblp_papers,intro)属性对,利用剩余的(部门,depart)、(姓名,name)及(coauthors,coauthor)属性对分别执行分块操作。这样就把问题转化为二分类问题,即:判断生成的候选表中的元组对是否为匹配元组对。

利用(姓名,name)属性对执行分块操作时,在候选表中产生的元组对数量最少。因此,从该候选表中选取约450条候选数据作为Sample数据集,再分别抽取Sample数据集中的10%、20%、30%、40%数据并把它们人工标注为训练集,然后将Tri-Training算法分别应用于K近邻分类器、随机森林分类器及决策树分类器,将小部分Sample数据集中的标签及其对应的特征纳入训练,最后比较不同大小的训练集和不同分类器下半监督学习算法获取标签的准确率。实验结果如表5所示。

可以看出,训练集中有标签的数据比例越大,Tri-Training算法的标签预测准确率越高。这意味着需要更多人工标注标签,因此折中选取了30%的有标签数据进行后续实验,并通过K近邻分类器来获取剩余数据的标签。在利用该方法完成剩余数据的标注后,就可以获得一个已标注的Sample标签数据集。

**表5 不同大小训练集及不同分类器下Tri-Training算法的标签获取准确率**

**Tab.5 Accuracy of Tri-Training for label acquisition under different size training sets and different classifiers**

| 分类器 | 不同数据标注的占比下获取标签的准确率 | | | |
|---|---|---|---|---|
| | 10% | 20% | 30% | 40% |
| K近邻 | 0.954 | 0.956 | 0.965 | 0.973 |
| 随机森林 | 0.952 | 0.955 | 0.961 | 0.962 |
| 决策树 | 0.943 | 0.950 | 0.971 | 0.970 |

将Sample标签数据集以7:3比例划分为训练集和测试集,并用其训练Magellan中提供的6种分类器,以找到最适用于此类数据的匹配器。以F1值作为评判标准,6种分类器在Sample测试集上的表现如表6所示。

**表6 Sample测试集在6种分类器上的表现**

**Tab.6 Performance of six classifiers on sample test set**

| 分类器 | 平均精度 | 平均召回率 | 平均F1值 |
|---|---|---|---|
| 决策树 | 0.959 | 0.989 | 0.974 |
| 随机森林 | 0.979 | 0.984 | 0.981 |
| 支持向量机 | 0.948 | 0.947 | 0.946 |
| 线性回归 | 0.946 | 0.954 | 0.948 |
| 逻辑回归 | 0.978 | 0.972 | 0.975 |
| 朴素贝叶斯 | 0.530 | 1.000 | 0.689 |

可以看出,随机森林分类器的表现较好,取得了最高的平均F1值。因此,执行论文数据集数据匹配时,选用随机森林作为分类器。

随后,需要选取合适的属性对,并对其进行搭配,以生成特征(即2个表中匹配属性对应的属性值)。实验中选取的特征如表7所示。其中,All表示采用所有的属性对;−(C1,C2)表示从All集合中去除(C1,C2)属性对;+(C1,C2)表示只采用该属性对生成特征。最终的数据匹配准确率结果如表7所示。

从表7可以看出,在用(姓名,name)匹配属性对执行分块操作以生成候选数据集时,无论采用何种特征属性对集合生成特征,其数据匹配的准确率都是最高的。因此,利用(姓名,name)匹配属性对生成候选数据集对结果最为有益。同时可以看出,只使用(部门,depart)属性对的匹配准确率高于使用全部属性对的匹配准确率,并且只使用(邮箱,email)匹配属性对生成特征时,匹配的准确率最高。此外,从

表7　在不同特征及不同分块操作生成的候选数据集下数据匹配准确率

Tab.7　Data matching accuracy under candidate datasets generated by different features and different blocking

| 特征属性对 | 分块选取属性对下数据匹配准确率/% | | |
|---|---|---|---|
| | (姓名,name) | (部门,depart) | (coauthors,coauthor) |
| All | 84.96 | 44.94 | 6.41 |
| −(部门,depart) | 37.67 | | 4.27 |
| −(姓名,name) | | 44.98 | 6.40 |
| −(职称,title) | 83.31 | 42.15 | 6.11 |
| −(学历,edu) | 84.58 | 44.32 | 6.26 |
| −(邮箱,email) | 85.09 | 44.97 | 6.38 |
| −(coatuhors,coauthor) | 85.57 | 44.28 | |
| −(dblp_papers,intro) | 88.60 | 45.03 | 6.78 |
| +(部门,depart) | 87.52 | | 6.84 |
| +(姓名,name) | | 24.86 | 3.74 |
| +(职称,title) | 12.37 | 4.68 | 1.32 |
| +(学历,edu) | 9.44 | 4.28 | 1.63 |
| +(邮箱,email) | 89.70 | 45.10 | 6.96 |
| +(coatuhors,coauthor) | 47.56 | 12.16 | |
| +(dblp_papers,intro) | 9.33 | 2.75 | 0.92 |

只使用(职称,title)、(学历,edu)或(dblp_papers,intro)属性对的匹配准确率可以看出,这三者对于匹配准确率的提升并无显著作用。原因在于,这3个属性对包含了较多空值,或包含了大量字符串数据,因此生成的相似度特征区分度不大。从表7可以得出如下结论:在利用(姓名,name)属性对对论文数据集执行分块操作,并且只使用(邮箱,email)属性来生成特征时,获得的数据匹配准确率最高。

### 2.3.2　视频元数据集数据匹配

首先,从第2.2.2节的视频元数据集模式匹配实验中,得到了名字、演员、导演、年份、地区、语言、标签共计7个公共属性。为了缩减候选实体对集的规模,从公共属性集中选取属性执行分块操作,排除因描述角度不同而在4个实验数据集中有较大语义差异的标签属性,以及空值较多的年份、地区、语言属性。根据名字、演员、导演3个属性的特性,分别应用不同的分块方法对生成的笛卡尔积进行筛选,具体如下:

(1)对名字属性应用Overlap Blocking。首先,用Trigram进行分词,并要求元组对至少在3个token上重叠,若名字属性值不足以分出3个token,则不进行过滤。

(2)对演员及导演属性应用Rule Based Blocking。过滤掉杰卡尔德相似度不足0.8的元组对,若演员或导演属性为空值,则不进行过滤。

经分块筛选后,得到候选实体对集。从候选实体对集中随机抽取500个实体对,进行人工标注。将人工标注后的500个实体对作为30%数据,并对其应用Tri-Training算法后,得到样本集$S$。随后,同样按照7:3的比例,把样本集$S$划分为训练集与测试集,再用6种分类器进行训练。最后,通过五折交叉验证找出最优分类器。训练集上五折交叉验证的实验结果如表8所示。

表8　6种分类器在训练集上的五折交叉验证结果

Tab.8　Five-fold cross-validation results of six classifiers on training set

| 分类器 | 平均精度 | 平均召回率 | 平均F1值 |
|---|---|---|---|
| 决策树 | 0.896 | 0.875 | 0.883 |
| 随机森林 | 0.920 | 0.929 | 0.924 |
| 支持向量机 | 0.836 | 0.818 | 0.826 |
| 朴素贝叶斯 | 0.918 | 0.890 | 0.904 |
| 线性回归 | 0.922 | 0.909 | 0.915 |
| 逻辑回归 | 0.938 | 0.914 | 0.925 |

从F1值可以看出,从训练集中抽取所有公共属性的特征时,随机森林分类器和逻辑回归分类器的分类效果较好。因此,在测试集上应用以上2种方法,并抽取不同特征进行对比实验。F1值如表9所示。表9中,All表示采用所有的公共属性,−A表示从All集合中去除A属性,+A表示只采用该公共属性生成特征。

从表9可以发现:随机森林与逻辑回归分类器的精度、召回率以及F1值都差别不大;在大部分情况下,随机森林分类器的表现略优于逻辑回归分类器。即便是同种方法,在各评估指标上进行纵向比较时仍然产生了较大差别。以F1值为例,在只利用地区、语言或标签这类区分度不高的属性生成特征时,F1值降低至80%以下。由此可见,数据匹配的

**表9 不同特征下在测试集上使用随机森林与逻辑回归分类器进行数据匹配的结果**

**Tab.9 Data matching results with random forest and logistic regression classifiers on test set under different characteristics**

| 特征 | 随机森林分类器 | | | 逻辑回归分类器 | | |
|---|---|---|---|---|---|---|
| | 精度 | 召回率 | F1值 | 精度 | 召回率 | F1值 |
| ALL | 0.965 | 0.944 | 0.954 | 0.951 | 0.938 | 0.944 |
| −名字 | 0.897 | 0.972 | 0.933 | 0.867 | 0.951 | 0.907 |
| −演员 | 0.972 | 0.958 | 0.965 | 0.932 | 0.958 | 0.945 |
| −导演 | 0.951 | 0.938 | 0.944 | 0.945 | 0.958 | 0.952 |
| −年份 | 0.934 | 0.889 | 0.911 | 0.924 | 0.924 | 0.924 |
| −地区 | 0.958 | 0.944 | 0.951 | 0.957 | 0.931 | 0.944 |
| −语言 | 0.965 | 0.951 | 0.958 | 0.951 | 0.938 | 0.944 |
| −标签 | 0.971 | 0.931 | 0.950 | 0.950 | 0.931 | 0.940 |
| +名字 | 0.916 | 0.910 | 0.913 | 0.935 | 0.903 | 0.919 |
| +演员 | 0.899 | 0.868 | 0.883 | 0.866 | 0.854 | 0.860 |
| +导演 | 0.941 | 0.889 | 0.914 | 0.866 | 0.896 | 0.881 |
| +年份 | 0.912 | 0.938 | 0.925 | 0.912 | 0.938 | 0.925 |
| +地区 | 0.696 | 0.889 | 0.781 | 0.667 | 0.917 | 0.772 |
| +语言 | 0.755 | 0.813 | 0.783 | 0.743 | 0.861 | 0.797 |
| +标签 | 0.747 | 0.799 | 0.772 | 0.788 | 0.799 | 0.793 |

效果主要取决于生成特征的质量,而非分类器本身的好坏。

### 2.4 数据融合实验

按照第1.4节中提出的数据融合流程进行实验。在数据匹配工作完成后,IR数据集与DBLP数据集已融合为一个数据集,3个来源不同的视频元数据集也被合并为一个数据集,只有公共属性尚未融合。依据第1.4节中的融合规则以及公共属性的数据特性,对名字、部门、职称等属性保留同一实体的较长属性值,对年份、语言、地区等属性保留同一实体的较多属性值,对演员、导演、标签等属性保留同一实体的属性并集。至此,多源数据集的实体已完成融合,但出现了一定数量的缺失值。缺失值的来源主要有以下2个方面:

(1)多源数据集本身的缺失。分以下3种情况:在某个公共属性上,多源数据集对含有缺失值的公共实体(对齐的实体)的描述均缺失;在某个非公共属性上,来源数据集对含有缺失值的公共实体的描述缺失;在实体融合前后,含有缺失值的非公共实体(未对齐的实体)在源数据集中的缺失属性形式相同。

(2)数据融合产生的缺失。原本不含缺失值的非公共实体,在数据融合后由于数据集属性列的扩充而产生新的属性缺失。

无论是原有的还是新产生的缺失,缺失信息都可能蕴藏在其他属性列的描述中。因此,为了减少融合后数据集的缺失,提升融合数据集的质量,对视频元数据集非公共属性中的DOUBAN_SCORE和TYPE以及公共属性中的语言和地区,进行缺失值填补。

由于DOUBAN_SCORE为数值型属性,采取均值、中位数、K近邻、支持向量回归、缺失森林、奇异值分解6种方法,对该属性进行填补。实验结果如表10所示:

**表10 6种方法对DOUBAN_SCORE属性列填补的结果**

**Tab.10 Results of filling the DOUBAN_SCORE attribute column for six methods**

| 方法 | 平均绝对误差 | 均方误差 | 均方根误差 | 平均绝对百分比误差 |
|---|---|---|---|---|
| 均值 | 1.097 | 2.052 | 1.432 | 0.199 |
| 中位数 | 1.083 | 2.061 | 1.436 | 0.202 |
| K近邻 | 0.972 | 1.694 | 1.302 | 0.169 |
| 支持向量回归 | 1.050 | 1.925 | 1.387 | 0.195 |
| 缺失森林 | 0.956 | 1.567 | 1.252 | 0.165 |
| 奇异值分解 | 1.215 | 2.359 | 1.536 | 0.206 |

可以看到,K近邻、支持向量回归、缺失森林3种方法中,缺失森林取得了最好的预测结果。奇异值分解方法利用矩阵补全的思想对数值型属性进行填补,取得的效果较差。这可能是由于实验数据集中数值型属性列较少,难以满足矩阵补全所需的低秩特征,而这种冗余性的缺少影响了奇异值分解方法的预测效果。

随后,按照第1.4节中描述的字符型缺失数据填补框架,对TYPE、语言和地区属性列进行填补。将类别属性建模为分类变量,先进行独热编码,再进行词嵌入来生成特征;将自由文本属性建模为序列变量,分别采用LSTM网络及*n*-gram hashing 2种方法完成实验,然后以众数填充作为基线方法完成对比。以准确率为评估指标,对实验结果进行评估。此处准确率定义为预测正确的样本占所有测试样本的比例。爱奇艺优酷腾讯元数据集、智能EPG补充元数据集、同洲媒资库元数据集以及融合后的数据集上的实验结果如表11所示。

从表11可以发现,LSTM网络及*n*-gram hashing在不同数据集的不同属性上各有优劣。总体而言,*n*-gram hashing的准确率稍好于LSTM网络。这可能是由于实验数据集中含有的长文本字段较少,LSTM网络在特征提取上的优越性难以发挥。同时对比3个属性的填补效果,TYPE属性上的填补准确率明显高于语言及地区属性,这与数据的分布特征一致。TYPE的种类少于其他2种属性,因此预测难度较小。此外,在语

**表 11　采用 3 种方法对不同数据集的不同属性列填补的准确率**

**Tab.11　Accuracy of imputation for different attribute columns in different datasets with three methods**

| 数据集 | 属性 | 各方法填补准确率/% | | |
|---|---|---|---|---|
| | | 众数填充 | LSTM 网络 | *n*-gram hashing |
| 爱奇艺优酷腾讯元数据集 | TYPE | 30.48 | 93.18 | 97.73 |
| | 语言 | 45.60 | 88.75 | 92.70 |
| | 地区 | 56.30 | 78.41 | 78.98 |
| 智能 EPG 补充元数据集 | 语言 | 43.53 | 89.58 | 89.20 |
| | 地区 | 31.12 | 86.81 | 81.25 |
| 同洲媒资库元数据集 | TYPE | 59.86 | 91.29 | 89.50 |
| | 语言 | 68.75 | 93.15 | 92.36 |
| | 地区 | 77.62 | 90.28 | 90.97 |
| 融合数据集 | TYPE | 22.56 | 94.36 | 98.33 |
| | 语言 | 44.92 | 90.24 | 92.20 |
| | 地区 | 50.29 | 90.63 | 90.05 |

言及地区属性方面，由于存在一些语义相同但表述不同的属性值，如“中国大陆”与“中国内地”，还有一些多值干扰，因此预测难度较大。总体来看，与众数填充方法相比，LSTM 网络及 *n*-gram hashing 的填充效果都有巨大提升，验证了所提出的缺失值填补框架的有效性。

### 2.5　有效性验证

基于半监督学习的多源异构数据治理框架的目标是：尽可能地减少人工参与，并尽可能提高数据治理过程中的自动化程度。因此，将数据治理过程中所花费的时间作为评估指标，在 5 个数据集上进行数据治理，然后对比全人工方式所花费的时间与所提出方法所花费的时间，如表 12 所示。其中，基于人工的数据治理，因不同个体所花费时间不一，所需时间基本为估计值；使用本研究中所提出的治理方法时，由于在数据匹配过程中仅需要少量人工标记数据，并且数据融合过程需要人工制定规则，因此这两部分所需的时间也为估计值。在数据匹配过程中，仅对少部分匹配数据做了人工标记，采用了半监督学习的方法进行预测，这在一定程度上导致匹配准确率有所降低，却也大大减少了人工数据匹配所需的时间。可以看出，在数据量较大的视频元数据集上，对比人工数据治理方法，本研究中提出的多源异构数据治理框架的效率大大提升。需要指出的是，有些任务几乎不可能依赖人工完成。

**表 12　利用 2 种方法进行数据治理所需时间对比**

**Tab.12　Comparison of time required for data governance between two methods**

| 数据集 | 步骤 | 人工数据治理所需时间 | 基于半监督学习的多源异构数据治理所需时间 |
|---|---|---|---|
| IR 数据集与 DBLP 数据集 | 模式匹配 | 约 30 s | 35 s |
| | 数据匹配 | 约 30～40 min | 约 421 s |
| | 数据融合 | 约 20 min | 约 10 min |
| 3 个视频元数据集 | 模式匹配 | 约 2 h | 2 min |
| | 数据匹配 | 10 年以上 | 约 1 h |
| | 数据融合 | 几乎不可能 | 约 8 h |

## 3　结语

设计并测试了一个基于半监督学习的多源异构数据治理架构，将现实世界中描述同一实体但来自不同数据源的异构数据整合为结构化数据。具体流程包括信息提取、模式匹配、数据匹配和数据融合 4 个部分。实验结果表明，该架构不仅能够有效破解“数据孤岛”状态，而且在尽可能减少人工参与的情况下显著提升数据质量。

**作者贡献声明：**

饶卫雄：数据治理方法提出，论文的撰写和修改。

高宏业：代码实现，实验验证，论文的撰写和修改。

林　程：代码实现，实验验证，论文的撰写和修改。

赵钦佩：方法和实验指导，论文的撰写和修改。

叶　丰：方法和实验指导，论文修改。

**参考文献：**

[1]　孟小峰，杜治娟. 大数据融合研究：问题与挑战[J]. 计算机研究与发展，2016，53(2)：231.
MENG Xiaofeng，DU Zhijuan. Research on the big data fusion：issues and challenges[J]. Journal of Computer Research and Development，2016，53(2)：231.

[2] DONG X L, SRIVASTAVA D. Big data integration[C]// 2013 IEEE 29th International Conference on Data Engineering (ICDE). New Jersey: IEEE Press, 2013: 1245-1248.

[3] SHVAIKO P, EUZENAT J. Ontology matching: state of the art and future challenges[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1): 158.

[4] TANG J, HONG M, ZHANG D L, *et al*. Information extraction: methodologies and applications [M]//Emerging Technologies of Text Mining: Techniques and Applications. New York: IGI Global, 2008:1-33.

[5] FINN A, KUSHMERICK N. Information extraction by convergent boundary classification[C]//AAAI-2004 Workshop on Adaptive Text Extraction and Mining (ATEM). San Jose: AI Access Foundation, 2004: 1-6.

[6] GHAHRAMANI Z, JORDAN M I. Factorial hidden Markov models[J]. Machine Learning, 1997, 29: 245.

[7] FREITAG D, MCCALLUM A. Information extraction with HMM structures learned by stochastic optimization [C]// Proceedings of the Sixteenth National Conference on Artificial Intelligence. Austin: AAAI Press, 2000: 584-589.

[8] LAFFERTY J, MCCALLUM A, PEREIR F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001: 282-289.

[9] CIRAVEGNA F. (LP)2, an adaptive algorithm for information extraction from web-related texts[C]//Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining, 17th International Joint Conference on Artificial Intelligence (IJCAI). Seattle: Morgan Kaufmann Publishers, 2001: 1-10.

[10] WANG X, KOMIYA J, SUHARA Y, *et al*. Koko: a system for scalable semantic querying of text[C]//Proceedings of the VLDB Endowment. Rio de Janeiro: VLDB Endowment, 2018: 2018-2021.

[11] BILENKO M Y. Learnable similarity functions and their application to record linkage and clustering[D]. Austin: The University of Texas at Austin, 2006.

[12] KOPCKE H, THOR A, RAHM E. Evaluation of entity resolution approaches on real-world match problems [J]. Proceedings of the VLDB Endowment, 2010, 3(1): 484.

[13] MUDGAL S, LI H, REKATSINAS T, *et al*. Deep learning for entity matching: a design space exploration [C]// Proceedings of the 2018 International Conference on Management of Data. New York: ACM, 2018: 19-34.

[14] TRIVEDI R, SISMAN B, MA J, *et al*. LinkNBed: multi-graph representation learning with entity linkage [C]// Proceedings of the 56th Annual Meeting of the Association for Computational. Melbourne: ACL, 2018: 252-262.

[15] KONDA P, ZHANG H, NAUGHTON J, *et al*. Technical perspective: toward building entity matching management systems[J]. ACM SIGMOD Record, 2018, 47(1): 33.

[16] BLUM A, MITCHELL T. Combining labeled and unlabeled data with co-training[C]//Proceedings of the Eleventh Annual Conference on Computational Learning Theory. New York: ACM, 1998: 92-100.

[17] CHEATHAM M, HITZLER P. String similarity metrics for ontology alignment[C]// Proceedings of the 12th International Semantic Web Conference. Berlin: Springer-Verlag, 2013: 294-309.

[18] YU L, CHEN S H, CHEN J. Property alignment of linked data based on similarity between functions [J]. International Journal of Database Theory and Application, 2015, 8(4): 191.

[19] 刘莎, 杨有龙. 基于灰色关联分析的类中心缺失值填补方法[J]. 四川大学学报(自然科学版), 2020, 57(5): 871.
LIU Sha, YANG Youlong. Imputing missing value by class center based on grey relational analysis[J]. Journal of Sichuan University(Natural Science Edition), 2020, 57(5): 871.

[20] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems [J]. IEEE Computer, 2009, 42(8): 30.

[21] ZHOU Z H, LI M. Tri-Training: exploiting unlabeled data using three classifiers [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(11): 1529.