基于特征挖掘的企业融资信息资源数据智能匹配方法

庞 泰[1] 翁 巍[1] 孟 灿[1] 赵 蕾[1] 牛红伟[1]
PANG Tai WENG Wei MENG Can ZHAO Lei NIU Hongwei

摘 要 企业融资信息数据多处于异构数据属性，在匹配过程中需要进行数据结构转换。由于现有数据匹配方法处理相似匹配特征，易受数据结构转换作用影响，导致智能匹配召回率偏低，因此，基于特征挖掘设计了一种企业融资信息资源数据智能匹配方法。提取了信息资源数据智能匹配特征，为有效区分异构数据属性，避免数据结构转换作用影响，进一步进行了信息资源智能匹配数据特征挖掘，保证数据智能匹配的准确性。最终基于挖掘后的特征，设计了跨模态信息资源数据智能匹配函数，从而实现了信息资源数据智能匹配。实验结果表明，设计的企业融资信息资源数据特征挖掘智能匹配方法，在不同类型融资信息资源数据下的智能匹配召回率均较高，匹配效果较好，具有准确性，有一定的应用价值，有利于推动企业信息化管理升级。

关键词 特征挖掘；企业；融资；信息；资源数据；智能匹配

doi：10.3969/j.issn.1672-9528.2024.03.038

0 引言

随着信息技术的不断发展，企业融资信息资源的数据量呈现出爆炸性增长，许多企业面临着融资难、融资贵的问题，这不仅限制了企业的发展，也影响了整个经济的活力[1-2]。在企业融资过程中，信息不对称是一个普遍存在的问题，导致融资效率低下和市场失灵[3-4]。同时企业融资信息资源的种类较多，数量庞大，传统的信息处理方式已经无法满足现代企业的需求，难以进行有效的数据匹配[5]。为了解决上述问题，需要设计一种可靠的企业融资信息资源数据智能匹配方法，以帮助企业更高效地获取和利用融资信息，降低融资难度，提高融资效率和成功率。

针对于此，相关研究领域人员已对数据匹配方法展开研究。例如，路璐等人[6]提出了多因子权重匹配方法，该方法利用距离、形状、大小、方向等空间相似性因子计算候选匹配对的各指标匹配概率，找出多种匹配关系，以完成匹配。但该方法用于企业融资信息资源数据匹配中，为充分考虑数据结构转换的影响，导致匹配结果存在偏差。陈源等人[7]利用自监督模型提取的文本数据对的交互信息，以特征增强的方式辅助基于神经网络的语义匹配模型，构建多任务的匹配模型。但该方法在企业融资信息资源数据匹配中，对于少数类别的匹配关系学习不充分，从而影响最终匹配结果的准确性。刘志丹等人[8]提出了一种空间结构匹配方法，利用同时满足查询关键字、距离和方向约束的空间对象构造对象连接图，基于多路连接的基准方法，将问题转换为在对象连接图上搜索与结构同构的子图匹配，并引入基于扫描线算法的边匹配计算，来降低匹配计算开销。但该方法在企业融资信息资源数据匹配中，未考虑不同维度属性之间的匹配关系，易受数据结构转换作用影响，导致匹配结果不理想。唐春兰等人[9]提出卷积神经网络的数据匹配方法。该方法首先预处理文本数据，然后利用主成分分析方法进行数据降维，提取数据特征，最后通过卷积神经网络算法实现文本数据匹配。但该方法用于企业融资信息资源数据匹配中，为充分考虑数据结构转换作用对匹配的影响，导致匹配召回率偏低。

企业融资信息数据多处于异构数据属性，在匹配过程中需要进行数据结构转换。由于现有数据匹配方法处理相似匹配特征，易受数据结构转换作用影响，导致智能匹配召回率偏低，因此，基于上述方法所存在的问题，为提高匹配效果，本文基于特征挖掘设计了一种全新的企业融资信息数据智能匹配方法。

1 企业融资信息资源数据特征挖掘智能匹配方法设计

1.1 提取信息资源数据可匹配特征

不同类型的企业融资信息数据的特征不同，为了提高企业融资信息的智能匹配效果，首先需要提取信息资源数据智能匹配特征，即确定信息资源数据智能匹配相似度，进一步对信息资源数据智能匹配问题进行描述，生成符合信息资源数据智能匹配特征提取流程，首先可以预设 A、B 两个属性

1. 青海省公共信用信息中心 青海西宁 810001

集合，计算 A、B 属性集合的相似度 $M_{A,B}$，其公式为：

$$M_{A,B} = Sim(A,B) \quad (1)$$

根据属性的相似度关系，可以将上述的属性集合与现有的数据库特征进行匹配，并进行特征筛选。本文设计的信息资源数据智能匹配方法根据初始匹配过程提取了数据智能匹配特征值，进行了相似度计算，此时根据匹配元数据状态可以进行双向过滤干扰，从而生成最终的特征提取映射，获得高精度数据智能匹配特征。该信息资源数据智能匹配特征提取流程主要包括预处理、分类、相似度计算、双向过滤干扰等四个基础步骤，该流程的具体组成如图 1 所示。

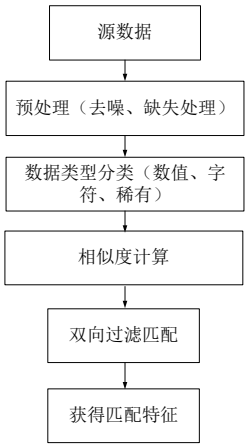

图 1 信息资源数据特征提取流程

由图 1 可知，在预处理阶段，需要对可能存在干扰的特征数据进行预处理，判断数据匹配列中存在的缺失数据的占比，删除不合理的匹配阈值。接下来，需要对连续变量进行统计，生成合理的智能匹配阈值；在数据类型分类阶段，需要根据数据的取值差异调整匹配标准；相似度计算阶段，需要利用注意力机制和神经网络进行数据训练，待其收敛后根据公式（1）计算相似度，记录数据相似度，判断信息资源数据的智能匹配关系。

待上述步骤完成后，本文使用双向过滤法获取了有效的数据特征参量，初步进行了匹配特征筛选，即使用源数据和待匹配的数据作为基础训练集，利用不同方向的神经网络进行相似性度量，建立匹配数据元素的语义映射关系，具体的筛选流程如下。

Step1：对源数据集合进行分类，提取对应数据类型的特征值，输出数值型特征向量。

Step2：将上述输出的特征向量作为基础输入值，训练神经网络。

Step3：将待匹配的数值特征向量输入到神经网络中，获取相似度矩阵，计算此时的数据匹配相似度。

Step4：设定合理的相似度阈值，输出满足相似度阈值的函数，生成正向匹配集合。

Step5：将上述生成的集合作为待匹配集合进行反相匹配，获取匹配交集，输出最终的数据智能匹配特征结果。

原始的数据智能匹配指标体系可能难以有效区分异构数据属性，获取的可利用信息不足，因此，本文设计的信息资源数据智能匹配方法需要根据数据模式信息与数据内容的差异性，调整数据限制关系，进一步进行数据匹配特征挖掘，保证数据智能匹配的准确性。

1.2 匹配数据特征挖掘算法设计

待上述的数据特征提取完成后，需要进一步进行数据特征挖掘，以有效区分异构数据属性，避免数据结构转换作用影响，提高后续匹配的准确性，即从数据中挖掘符合匹配要求的信息特征，判断其与提取特征之间的关系，获取数据背后的规律模式，提高数据智能匹配的准确性。基于此，进行信息资源数据特征挖掘，其挖掘流程如图 2 所示。

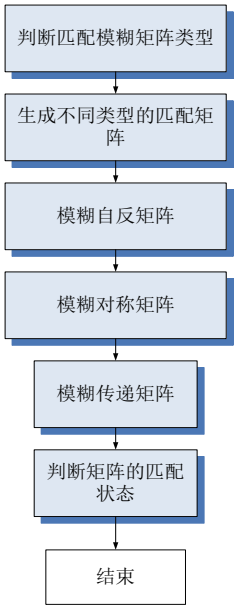

图 2 信息资源智能匹配数据特征挖掘流程

由图 2 可知，部分信息资源数据属于非结构化数据，具有复杂性和多样性，难以进行针对性描述，只能通过物理或逻辑层面来表达，因此，在数据挖掘的过程中，需要按照特征的分散关系依次进行处理，具体的数据特征挖掘流程如图 3 所示。

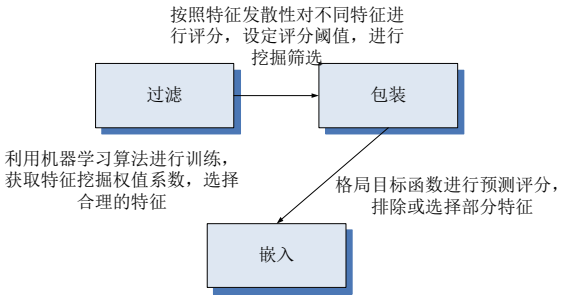

图 3 信息资源智能匹配数据特征挖掘流程

在实际特征挖掘的过程中，受信息资源数据的粒度不同的影响，可能出现特征多样性问题，需要根据数据对象的状态进行独立挖掘，保证挖掘的针对性。当待匹配的信息资源数据处于相同的汉明空间中时，还可能会出现感知异构问题，可以预先判定数据的哈希码，生成最优映射函数，完成异构数据关系感知。此时可以根据数据抽象关系进行特征挖掘，生成不同类型的匹配功能库，即运用特征挖掘法获取差异化数据特征，进行特征类型分类，再将原始的数据特征作为依据，进行数据信息存储，生成数据特征智能匹配索引，最后再进行匹配接口查询，采用差异化查询对策对匹配关系数据库进行对比。为了接收非结构对象，明确智能匹配对象的类型，有效地进行特征存储与查询，本文还使用了 CL 数据智能匹配处理引擎。该处理引擎可以对待匹配的数据特征进行自主分析，建立不同类型的数据特征对象，从而构建全新的数据挖掘索引，完整地执行数据智能挖掘过程。

在特征挖掘过程中，若出现了挖掘过拟合问题，就需要重新进行挖掘处理，即根据数据集之间的特征共性假定数据的唯一关系，此时不同匹配数据的度量空间不同，需要满足的数据匹配条件不同，因此，基于特征挖掘对提取出的特征进行分析和选择，选择出与融资相关性较强的特征，去除无关或冗余的特征。再进行数据特征挖掘流程中的模糊运算，生成不同类型的智能特征匹配数据库。这些数据库中每一个数据对应的特征不同，可以快速进行查询、检索，此时挖掘后的数据特征结构 $O(B,S)$ 为：

$$O(B,S)=\left\|X-B_S\right\|^2+\sum\lambda\left\|s_i\right\| \tag{2}$$

式中：X 代表初步提取获得的匹配数据特征，B_S 代表潜在匹配参量，λ 代表子空间映射系数，s_i 代表损失常数。根据上述挖掘的数据特征结构可以生成跨模态数据关联式，得到信息资源数据智能匹配函数，以完成企业融资信息资源数据智能匹配，具体描述如下。

1.3 设计跨模态信息资源数据智能匹配函数

在企业融资信息资源数据智能匹配的过程中，可能受语义模态变化影响，导致匹配检索异常，降低最终的智能匹配效率，因此，本文基于上述挖掘后的企业融资信息资源数据特征，设计了跨模态信息资源数据智能匹配函数，有效处理了语义标签，此时生成的数据智能匹配特征集合 ψ 为：

$$\psi=\frac{(\mathrm{v}_t,t_i)^n}{O(B,S)} \tag{3}$$

式中：V_t 代表资源数据样本特征向量，t_i 代表匹配文本样本特征向量，n 代表待匹配的数据信息资源数量。基于上述匹配集合可以快速获取语义标签量，确定特征匹配的维度 [10]。

信息资源数据智能匹配过程也可以看成跨模态任务检索过程，即从相应的数据库中查询符合匹配特征要求的样本 [11]，因此，需要由特征提取中心和子空间映射中心组成。经过预训练的数据智能匹配样本的特征可能存在一定的差异，难以有效地进行匹配标注 [12]，因此，可以根据多模态训练原理，有效地处理同现信息、成对信息，进行信息排列 [13]，赋予信息资源数据匹配标签，快速进行匹配区分。事实上，不同信息资源数据特征往往通过不同的方式提取，因此，需要对不同的信息资源数据进行相似性度量，计算的跨模态数据关联式 max 为：

$$\max=\frac{w_x\sum_{xy}w_y}{\sqrt{w_x\sum_{xx}w_x}\sqrt{w_y\sum_{yy}w_y}} \tag{4}$$

式中：w_x、w_y 分别不同的模态学习映射，$\sum_{xx}$、$\sum_{yy}$ 代表模态数据的协方差矩阵。根据信息数据资源的关联性可以构建全新的匹配语义空间，执行跨模态检索 [14]，此时部分典型数据匹配分析结构如图 4 所示。

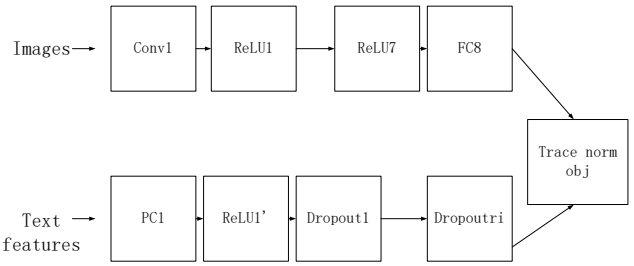

图 4 典型数据匹配分析结构

由图 4 可知，上述的典型数据匹配分析结构具有深度分析特征，因此，可以利用深度神经网络融合信息资源数据智能匹配特征 [15]，此时的代价函数 $corr(X, Y)$ 为：

$$corr(X,Y)=t_r(\frac{T^TT}{\psi})^{1\backslash2} \tag{5}$$

式中：T^T 代表正向匹配特征，T 代表反向匹配特征，t_r 代表梯度处理参量。根据上述的代价函数，可以将不同模态的信息资源数据映射到子空间中，计算数据特征的欧式距离，从而得到合适的匹配超参数，获取成对的匹配监督信息，剔除数据智能匹配集合 ψ 中语义不匹配的资源数据样本。

在上述的基础上，为有效地处理信息特征排列问题，赋予信息资源数据特征匹配标签，快速进行匹配区分。接下来，通过损失函数来进行多模态训练，此时的损失函数 $\arg\min L$ 为：

$$\arg\min L=\frac{\mu_{\max}}{2}\left\|w\right\|^2+\left\|w_q-w_v\right\| \tag{6}$$

式中：$\mu_{\max}$ 代表待匹配相似度，w 代表间隔排序损失，w_q、w_v 分别代表不同的相关变换矩阵，根据上述的损失函数可以有效衡量样本的关系，通过文本映射空间向量获取描述数据 [16]，从而提取数据智能匹配表征，获得新的数据智能匹配特征集合 X_ψ。

由于，信息资源数据智能匹配数据之间存在较大的分布式差异，为提高匹配精度，在上述数据智能匹配特征集合 X_ψ 中进行线性搜索，通过二值化表征学习对其匹配数据进行相似度比较，基于此，设计的跨模态信息资源数据智能匹配函数 $D(a)$ 为：

$$D(a)=\sum_{k=1}^{X_\psi} d(y_i^{(k)}, y_j^{(k)}) \tag{7}$$

式中：d 代表汉明距离，$y_i^{(k)}$ 代表企业融资信息资源模态数据，$y_j^{(k)}$ 代表数据智能匹配特征集合 X_ψ 中的数据语义表征。依据式（7）计算结果，当其得数值大于等于 1 时，说明其企业融资信息资源数据对相匹配，输出其匹配结果，完成企业融资信息资源数据智能匹配。

2 实验

为了验证设计的基于特征挖掘的企业融资信息资源数据智能匹配方法的匹配效果，本文配置了基础实验环境，选取了可靠的实验平台，将其与文献 [8] 和文献 [9] 中两种常规的企业融资信息资源数据智能匹配方法对比，进行了实验。

2.1 实验准备

结合企业融资信息资源数据智能匹配实验要求，本文选取 ModelWhale 云端数据分析平台作为实验平台，该实验平台能创建不同类型的实验数据源，包括数据集、存储连接对象、NAS 空间等，还可以快速标注实验数据，降低数据智能匹配实验难度。本文从 Wikipedia、MSCOCO 中抽取实验数据集，进行实验数据重组。初步获取的实验数据集中的数据格式不一，个别数据还包含语义标签，难以进行智能匹配，因此，在实验前，本文使用 VGG 深度神经网络进行了数据训练，提取了符合实验要求的文本特征。

针对实验要求，本文选取 MVC 作为基础模式，开发了 B/S 实验架构，如图 5 所示。

图 5 B/S 实验架构

在实验前，需要根据智能匹配交互状态利用 JavaScript 等编写数据匹配规则，避免出现严重的实验偏差。

2.2 指标设置

待上述的实验准备完成后根据数据智能匹配要求，本文选择数据智能匹配召回率 R_Z 和覆盖率 P_Z 作为实验指标，两指标的计算式为：

$$R_Z=\frac{a_i}{c_i}\times 100\% \tag{8}$$

$$P_Z=\frac{p_i}{p_j}\times 100\% \tag{9}$$

式中：a_i 代表正确匹配的第 i 个类别信息资源数据量，c_i 代表信息资源数据总量。数据智能匹配召回率越高证明数据智能匹配的效果越好，反之则证明数据智能匹配的效果相对较差。p_i 表示匹配到的相关数据数量，p_j 表示全部相关数据数量。覆盖率越高表示匹配方法能够较为全面地找到相关数据，相对来说更加有效。反之则意味着匹配方法可能存在漏匹配的情况，未能找到部分相关数据。

2.3 实验结果与讨论

在预设的实验环境下，选取若干个不同类别的实验信息资源数据集（资产负债、利润、现金流量、财务比率、销售额、市场份额、客户结构、供应链管理、市场规模、行业竞争、行业政策、担保、抵押物、评估价值、证券投入，其编号依次为 1 ～ 15），此时，分别使用本文设计的基于特征挖掘的企业融资信息资源数据智能匹配方法、文献 [8] 的空间结构匹配方法，以及文献 [9] 的卷积神经网络匹配方法进行数据智能匹配，使用公式（8）计算三种方法的数据智能匹配召回率，实验结果如图 6 所示。

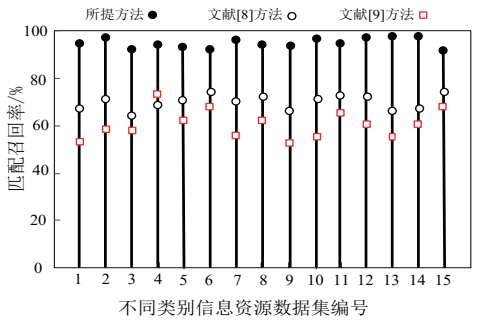

图 6 匹配召回率结果

由图 6 可知，本文设计的基于特征挖掘的企业融资信息资源数据智能匹配方法在不同融资信息资源数据类型下的数据智能匹配召回率均较高，文献 [8] 的空间结构匹配方法，以及文献 [9] 的卷积神经网络匹配方法进行数据智能匹配在不同融资信息资源数据类型下的数据智能匹配召回率相对较低。由此表明，本文方法的匹配效果较好，具有可靠性，有一定的应用价值。

接着，针对覆盖率指标，在上述测试背景下对上述方法展开对比测试，根据公式（9）计算出三种方法的覆盖率，其结果如图 7 所示。

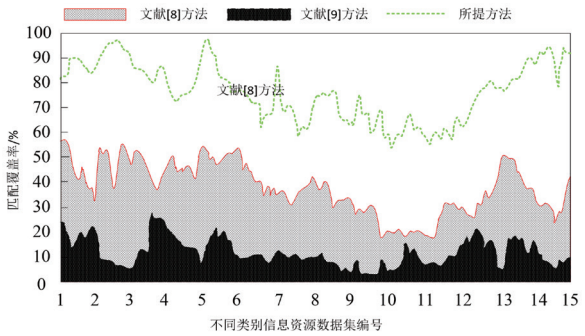

图 7 匹配覆盖率结果

由图 7 可知，本文设计的基于特征挖掘的企业融资信息资源数据智能匹配方法在不同融资信息资源数据类型下的数据智能匹配覆盖率均较高，最低均超过 60%；文献 [8] 的空间结构匹配方法，以及文献 [9] 的卷积神经网络匹配方法进行数据智能匹配在不同融资信息资源数据类型下的数据智能匹配覆盖率相对较低，均在 60% 以下。对比三种方法所得结果，得出本文方法能够较为全面地找到相关数据，更加有效。

3 结语

综上所述，随着大数据技术的不断发展，越来越多的企业和机构开始利用大数据技术来提高工作效率和准确性。在企业融资领域，通过大数据技术可以对海量的融资信息进行快速、准确的分析和处理，从而为投资者提供更加全面、准确的投资决策依据。受数据的种类及来源影响，目前的信息处理难度较高，很多常规的信息资源智能匹配处理技术需要使用 K-NN 算法获取匹配特征，极易受数据结构转换作用影响，导致匹配效果较差，因此，本文基于特征挖掘设计了一种全新的企业融资信息数据智能匹配方法。实验结果表明，本文设计的基于特征挖掘的企业融资信息资源数据智能匹配方法的匹配效果较好，具有可靠性，有一定的应用价值，有助于提高企业综合融资实力。

参考文献：

[1] 徐嘉硕，祁凯．突发公共卫生事件微博网络社群挖掘及特征研究：基于 2022 年上海疫情事件的分析 [J]. 情报探索，2023(11): 74-80.

[2] 孔晨晨，张沛，黄瑛，等．基于目标群体指数的农村地区交通事故汽车通行隐患特征挖掘方法 [J]. 道路交通管理，2023(10): 36-39.

[3] 王雪纯，毛华松，吴映华夏．基于古诗词文本挖掘的唐宋三峡人文景观特征及审美认知研究 [J]. 热带地理，2023, 43(10): 2001-2011.

[4] 罗敏，杨景旭，周尚礼，等．面向产业链特征挖掘和有序用电行业筛选的行业关联图模型 [J]. 电器与能效管理技术，2023(8): 53-60.

[5] 彭贤哲，周海玲，石进．图书馆服务场景下中文图书被引特征的挖掘、分析与应用：以 G 类图书为例 [J]. 现代情报，2023, 43(10):107-119.

[6] 路璐，孟妮娜．不同比例尺居民地数据的多因子加权匹配方法 [J]. 甘肃科学学报 ,2021,33(2):33-37.

[7] 陈源，丘心颖．结合自监督学习的多任务文本语义匹配方法 [J]. 北京大学学报 (自然科学版), 2022,58(1):83-90

[8] 刘志丹，林维鑫，伍楷舜．一种面向大规模空间文本数据的空间结构匹配算法 [J]. 计算机学报 ,2022,45(6):1261-1275.

[9] 唐春兰．基于卷积神经网络的近红外光谱文本数据匹配检测方法 [J]. 激光杂志 ,2022,43(10):106-110.

[10] 蒋浩杰，刘中国，丁国勇．近五年教育数据挖掘研究的特征：基于 184 篇 EDM 国际会议短论文的分析 [J]. 信息技术与信息化，2023(4):15-20.

[11] 吴若航，储节旺．知识付费在线课程试用者转移行为需求特征与负面影响因素挖掘研究：以“哔哩哔哩课堂”为例 [J]. 情报科学 ,2023,41(4):17-25.

[12] 陈勇，邹皓天，苏剑，等．基于数据挖掘的土水特征曲线影响分析及预测研究 [J]. 应用基础与工程科学学报，2023, 31(2): 451-466.

[13] 张广海，袁洪英．非物质文化遗产资源时空分布特征及景区响应机制：基于网络文本信息挖掘与定量测度分析 [J]. 地域研究与开发 ,2023,42(1):102-107.

[14] 周晓春．中国英语笔译能力等级量表描述语特征分析：基于文本挖掘的方法 [J]. 安徽工业大学学报 (社会科学版), 2022, 39(3): 43-47.

[15] 安相丞，陈蓉晖．我国高校师德问责制相关政策文本的基本特征研究：基于 ROST 数据挖掘系统的分析 [J]. 江汉大学学报 (社会科学版), 2022,39(3):99-109+127-128.

[16] 杨波，罗时达，刘韫尔．重大突发公共危机事件背景下客户特征比较研究：基于大数据挖掘方法 [J]. 江苏社会科学，2022(3): 146-155.

【作者简介】

庞泰（1974—），男，甘肃会宁人，硕士，高级工程师，研究方向：计算机、金融、社会信用。

翁巍（1975—），男，安徽六安人，本科，工程师，研究方向：计算机、金融、社会信用。

孟灿（1989—），男，安徽宿松人，硕士，高级工程师，研究方向：计算机、金融、社会信用。

赵蕾（1991—），女，藏族，青海西宁人，硕士，助理工程师，研究方向：计算机、金融。

牛红伟（1986—），男，河南新乡人，硕士，高级工程师，研究方向：计算机、金融、社会信用。

（收稿日期：2023-12-18）