文章编号：1001-4632（2026）02-0232-12
引用格式：郝晓培，阎志远，张军锋，等. 基于数据知识库的铁路客票敏感数据智能识别技术应用研究[J]. 中国铁道科学，2026，47(2)：232-243.
Citation：HAO Xiaopei，YAN Zhiyuan，ZHANG Junfeng，et al. Research on the Application of Intelligent Recognition Technology for Sensitive Railway Ticket Data Based on Data Knowledge Base [J]. China Railway Science，2026，47 (2)：232-243.

基于数据知识库的铁路客票敏感数据智能识别技术应用研究

郝晓培，阎志远，张军锋，李 雯，刘相坤，石瑞君

（中国铁道科学研究院集团有限公司 电子计算技术研究所，北京 100081）

摘 要：为应对铁路客运数据规模激增衍生的数据安全风险，实现敏感信息智能识别与动态防护，提出基于数据知识库的铁路客票敏感数据智能识别技术。通过构建“法律法规—行业标准—企业规范”3级知识库，结合铁路客票历史数据，设计多层次敏感数据智能识别算法，实现对多模态数据中敏感信息的高效精准识别。在此基础上引入图技术，构建数据资产及敏感数据血缘关系图谱，依据数据间流转拓扑关系，完成敏感信息标签在相关数据节点间的高效传播。结果表明：所提技术在结构化数据处理方面，可实现约21.7万条·s^{-1}的敏感信息识别效率，约为传统方案的2倍；在非结构化数据处理方面，通过领域知识图谱注入，将敏感实体识别的F_1值提升至91.24%，上下文误判率下降至5.88%；多媒体图片文本提取及敏感信息识别准确率达93.71%。该技术可显著提升铁路客票敏感数据识别的准确性与处理效率。

关键词：敏感数据；知识库；铁路客票；智能识别；标签传播；血缘关系图谱
中图分类号：U293.22 **文献标识码**：A **doi**：10.3969/j.issn.1001-4632.2026.02.20

铁路客票系统作为国家关键信息基础设施[1]，在为旅客提供出行服务的同时，也承担着保护海量旅客数据安全的责任。近年来铁路客运发送量持续攀升，2025年全国铁路旅客发送量超过45亿，日售票量超过2 000万，客票数据规模也随之急速膨胀，涵盖旅客的个人基本信息、出行信息、支付信息等。这些数据对铁路优化运营、提升服务质量、产品设计及推进供给侧改革意义重大，且随着我国数据要素政策的不断完善，其有效利用与流通可提升运输效率、促进多式联运与跨产业协同发展。然而，铁路客票数据中包含大量敏感信息，一旦泄露不仅会导致旅客个人信息被滥用，造成严重的隐私侵害，还可能引发社会公众对铁路部门的信任危机，对铁路声誉产生负面影响，从而影响铁路客运的可持续发展。当前各国纷纷加强了对数据安全和隐私保护的立法和监管力度，欧盟出台的《通用数据保护条例》对数据管理者的主体责任、安全措施及泄露响应等方面设置了严格的标准；我国也相继颁布《网络安全法》《数据安全法》《个人信息保护法》等一系列法律法规，明确了数据处理中的安全责任与合规要求，对铁路客票系统敏感数据的保护提出了更高的要求。在此背景下，高效准确地识别客票系统中文本及多媒体类数据的敏感信息，是实现数据资产化、保障敏感信息安全脱敏、实施差异化分级管控以及在数据要素流通中释放价值的坚实基础。

在文本类数据敏感信息识别方面，已有较多相关研究。符泽凡等[2]提出了基于双向编码器表示

收稿日期：2025-06-24；修订日期：2026-03-12
基金项目：中国铁道科学研究院集团有限公司院基金课题（2024YJ228）
第一作者：郝晓培（1990—），男，河南林州人，助理研究员。E-mail：linuxstar@126.com
通讯作者：李 雯（1987—），男，湖南邵阳人，副研究员。E-mail：1556860343@qq.com

模型（Bidirectional Encoder Representations from Transformers，BERT），结合变体字还原算法的网站敏感信息识别方法，为网页的内容敏感信息识别提供了新思路。李扬等[3]通过定义色情、暴力、违禁、邪教、反动等5类共2 639个敏感关键词，构建了敏感关键词与情感极性协同分析的敏感信息识别方法。Li等[4]从敏感信息特征中提取出支持向量，对支持向量机（Support Vector Machine，SVM）进行训练，提高了网络敏感信息识别的检测速度与准确性。Xu等[5−6]将主题聚类融入敏感信息识别中，通过构建基于加权潜在狄利克雷分布的网络敏感信息主题识别方法，将所得主题信息与经过双向循环神经网络（Bidirectional Recurrent Neural Network，Bi-RNN）表征后的文本特征向量进行融合，利用注意力机制进行权重计算实现敏感信息识别。

图像中敏感文字信息检测的研究方法主要可分为2类：基于传统视觉特征的敏感信息检测和基于图像文本特征的敏感信息检测。Krasser等[7]重点考虑图像的边缘方向一致性矢量、尺度不变特征变换特征以及颜色直方图，将这些视觉特征作为线性支持向量机分类器的判别标准。Wang等[8]将线分布云与最大稳定极值区域算法相结合以提取文本区域，该方法在弱光条件下仍具有较好的检测效果。随着光学字符识别（Optical Character Recognition，OCR）技术的不断发展，CRNN算法[9]和基于注意力机制的文本识别算法已成为当前主流。CRNN算法先通过卷积神经网络提取图像的空间特征，再利用循环神经网络将空间特征转化为序列特征，最终输出对应的文本内容，从而提升识别性能；基于注意力机制的识别算法通过引入注意力权重矩阵，根据输入序列的特征动态计算每个元素的重要性，实现更精准的文字识别。

本文提出基于数据知识库的铁路客票敏感数据智能识别技术，基于铁路领域行业词典以及数据安全法律法规，通过法律条文语义解析，并与铁路客运业务规则形式化映射，形成“法律法规—行业标准—企业规范”3层关联的知识库。针对不同类型的数据，综合考虑效率及准确率，设计多模态敏感信息混合识别引擎，针对结构化数据设计正则表达式与改进型Aho-Corasick自动机的双模式匹配算法，对非结构化文本采用知识增强型BERT模型实现上下文敏感语义感知，多媒体图像数据通过OCR-Transform模型提取文本信息从而实现敏感信息识别，最后基于数据血缘关系的敏感数据标签传播，形成敏感数据在跨系统流转中的拓扑关系与权限演变路径。从而解决铁路数据安全治理中合规落地难、识别精度低、跨域溯源能力弱等问题。

1　铁路客运敏感信息识别体系框架

为满足铁路客运多模态数据在跨场景流转中的安全合规要求，基于“知识驱动—智能识别—血缘关系”识别体系，通过构建动态铁路知识库，研发“规则＋AI”协同多层次敏感信息智能识别模型体系，并结合数据血缘关系，实现敏感信息从智能识别到流转、处理的全链路高效传播。铁路客运敏感信息识别体系框架如图1所示。首先，利用本体建模与语义解析技术，将国家法律法规、行业规范及企业内部治理标准转化为结构化知识库，该知识库包含敏感信息种类、数据分类分级、铁路客运术语

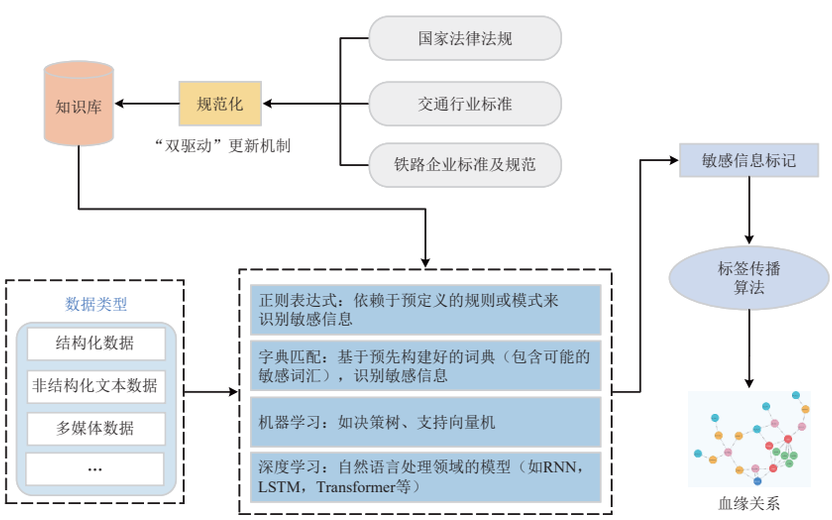

图1　铁路客运敏感信息识别体系框架

等核心逻辑，通过“双驱动”（定时＋手动）更新机制实现知识库的动态更新，以支持合规信息的多维度检索。其次，铁路客运数据包含结构化数据、非结构化文本及多媒体图片等多模态类型，需针对不同数据设计相应的敏感信息识别模型。其中，结构化数据采用正则表达式与预定义敏感字段库，实现毫秒级高效精准识别；非结构化文本通过BERT模型捕捉上下文隐含的敏感信息；多媒体图片则结合目标检测与OCR文字提取技术，自动识别并提取其中的文本信息。最后，通过自动化采集并解析数据在采集、存储、整合、使用与呈现、分析与应用、归档和销毁等全生命周期各环节的元数据与操作日志，构建敏感数据流转关系图谱，并基于知识库的规则信息实现敏感数据标签的动态传播。

综上，通过建立知识库与多模态智能识别引擎的动态协同机制，可有效弥补传统规则引擎的覆盖盲区。依托数据血缘分析技术实现客票敏感数据跨系统的数据流转拓扑，将原本分散的敏感数据管理整合为有机管理网络，形成1套兼顾安全与效率的标准化敏感信息识别框架。

2 铁路知识库

铁路客票数据知识库以各类法律、标准及规范为支撑构建。其中，法律法规涵盖《网络安全法》《数据安全法》《个人信息保护法》；行业标准包含《交通运输数据安全风险评估指南》；企业规范包括《铁路旅客运输规程》《铁路个人信息保护标准》等。利用DeepSeek大模型对相关文件进行实体抽取，如“重要数据”“个人敏感信息”“出行记录”等关键词并进行关系标注，将条款语义结构化处理，形成可供模型直接读取的知识库列表。通过“人工采集＋实时监测爬取”双驱动更新机制，自动获取法律法规发布平台的动态信息，并经合规专家审核后更新入库，确保知识库的时效性与准确性。其技术框架如图2所示。

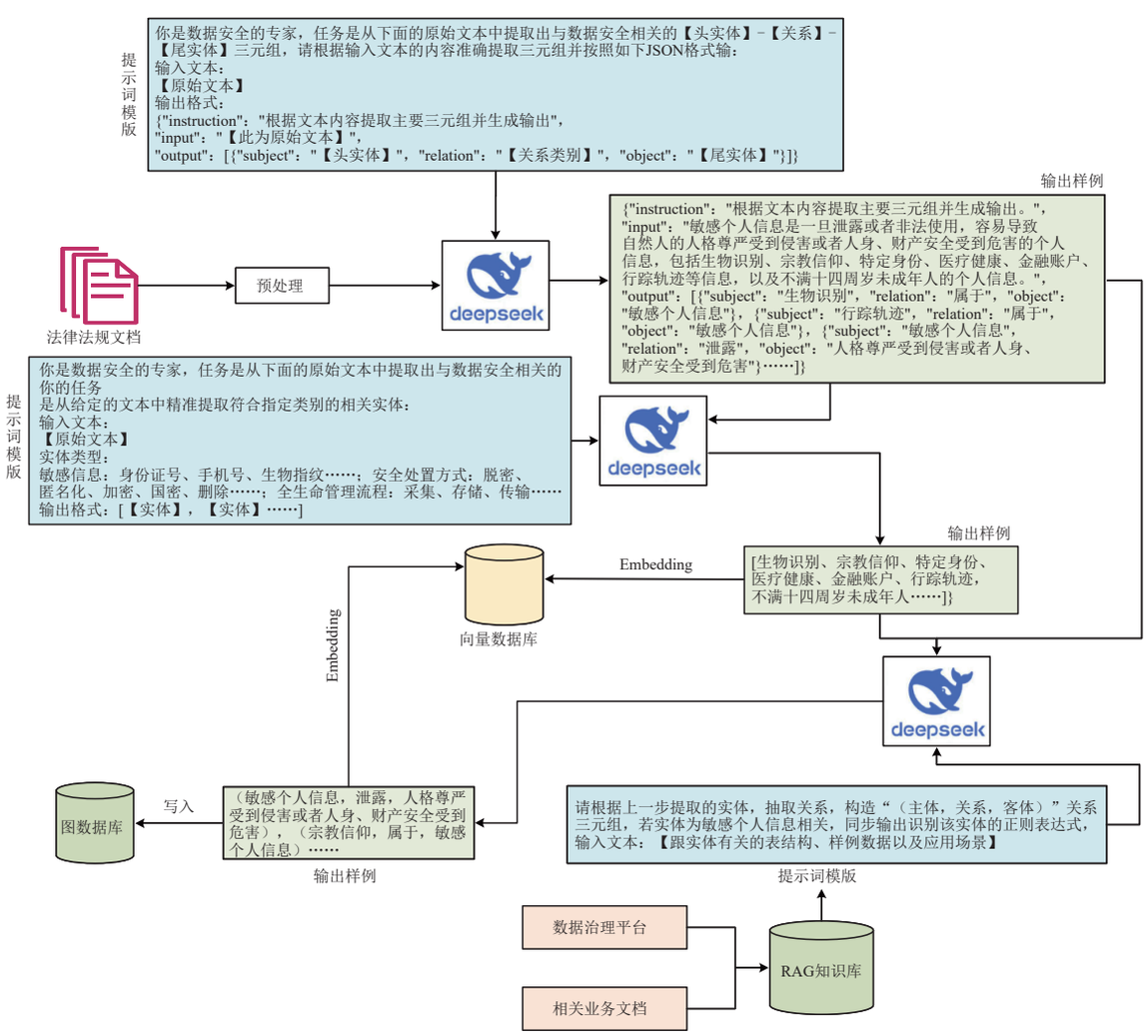

图2 铁路敏感信息知识库构建技术框架

相关要求主要以非结构化文本数据为主，首先，对文档材料进行预处理，划分为若干具有明确语义边界的语义块。随后，引入具有强大语言理解与推理能力的DeepSeek大语言模型，以原始文档的语义块及定制的提示词模板作为输入，按照JSON格式输出提取的核心信息，并将其解析为结构化数据，完成语义信息的结构化存储。最后，对结构化语义块进行逐条解析，采用基于向量检索与LLM一致性判定的增量机制，实现知识库的动态扩展。该机制的核心流程包括3步：①语义提取，基于DeepSeek与定制提示词，从语义块中提取JSON格式的目标信息（实体或关系）；②向量表征，对提取结果进行Embedding操作，得到向量表示；③双重校验，通过向量匹配（余弦相似度）初筛重复项，再经LLM一致性判定语进行确认，确保仅新增信息并入全局集合。

基于上述流程，该机制分别实现了实体与关系的动态维护。在实体维护中，初始化阶段对首个文档提取初始实体集，经Embedding后写入向量数据库作为全量实体集；后续新增文档按相同方式提取增量实体集，经双重校验后仅将新增实体并入，实现动态扩展。在关系维护中，将JSON格式的语义块与全局实体集共同作为DeepSeek输入，结合定制提示词完成三元组抽取。初始化阶段生成的三元组作为全局关系基础写入向量数据库；后续每个语义块提取的增量局部三元组，经双重校验后自动归并重复关系、扩展关系类型体系，最终写入图数据库，形成合规知识图谱。

铁路敏感信息知识库示例如图3所示。以身份证号为例，基于知识库可以确定其属于敏感个人信息，从而完整展现身份证号在客票系统敏感个人信息保护语境下的处理逻辑、识别规则等。基于知识库生成的敏感信息实体见表1，同时依照合规要求，按相关实体在法律法规中出现的频次及重要程度动态维护实体权重。

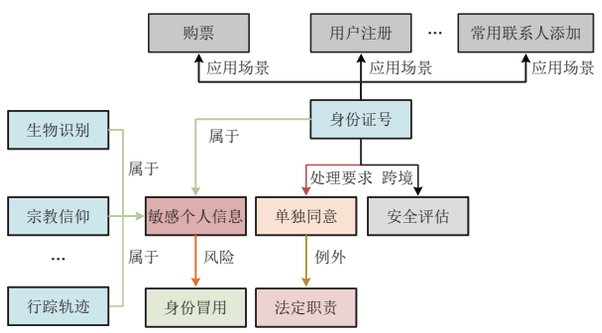

图3　铁路敏感信息知识库示例

表1　敏感信息实体（部分）

名称	依据	特征描述	典型示例	敏感程度
身份证号	《个人信息保护法》第4条	唯一绑定自然人身份，泄露可导致诈骗、身份冒用	110105******123X	1.0
手机号	《网络安全法》第41条	结合其他信息可定位个人，易引发骚扰电话、钓鱼攻击	138****1111	0.9
出行记录	《数据安全法》第3条	反映个人行为轨迹，泄露可能威胁人身安全	20250423高铁G101次北京→上海	1.0
支付信息	《网络安全法》第21条	直接关联资金安全，泄露可能导致财产损失	银行卡号6217****1234	1.0
联系地址	《个人信息保护法》第6条	住址泄露易引发人身安全风险	北京市海淀区**路	0.8
未成年人信息	《个人信息保护法》第29条	不满14周岁人群信息受特殊保护，泄露危害更严重	儿童出生日期、监护人联系方式	1.0
违禁词库	法律法规/社会文化	暴恐违禁、文本色情、政治敏感、低俗辱骂等		1.0

3　多模态敏感信息混合识别引擎

多模态敏感信息混合识别引擎主要识别客票系统结构化数据、非结构化文本和多媒体图像数据中的敏感信息。针对不同模态的数据特点，设计相应的识别算法，该引擎的数据处理链路如图4所示。分为模型训练和敏感信息识别2个阶段，在模型训练阶段，先对客票系统内数据进行预处理及分类，

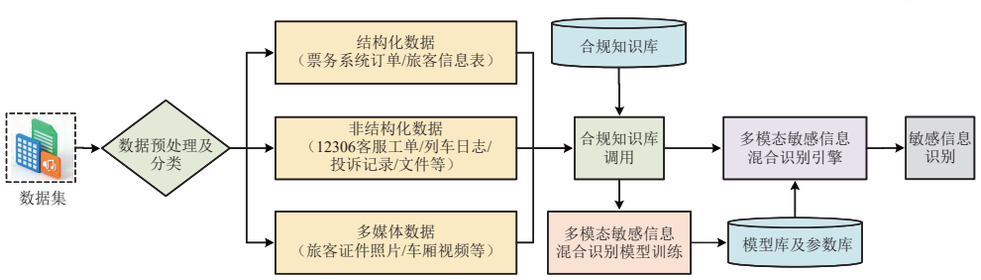

图4　多模态敏感信息混合识别引擎数据处理链路

调用知识库获取敏感信息实体（包括：内容、样例、识别规则等），再将预处理后的数据与敏感信息实体信息输入相应的识别算法进行训练，得到针对每种敏感信息实体在不同数据类型上的识别模型，并将模型参数存入数据库，供实时识别阶段调用；在敏感信息识别阶段，引擎加载训练好的模型，对新产生的多模态数据进行实时检测，输出敏感信息检测结果。

3.1 结构化数据敏感信息识别算法设计

目前客票系统包括注册、常用联系人及行程轨迹等敏感信息，这些存储在数据库和业务日志中，通过正则表达以及改进型 AC 自动机算法进行敏感信息识别。

1）正则表达式

正则表达式通过普通字符与元字符组合定义文本匹配规则，可实现身份证号、手机号、电子邮箱等结构化敏感字段的识别，其优势在于规则定义灵活、无须复杂模型训练，但也存在计算效率随数据规模非线性下降、语义泛化能力不足及维护成本高等局限，更适用于小规模、规则明确的场景。典型敏感信息字段及正则表达式见表 2。

表 2 典型敏感信息字段及正则表达式

字段类型	正则表达式
身份证号	/^[1-6]\d{5} (18|19|20)\d{2} (0[1-9]|1[0-2]) (0[1-9]|[12]\d|3[01])\d{3}[\dXx]$/
手机号	^(?:(?:\+|00)86)?1(?:3\d|4[5-79]|5[0-35-9]|6[5-7]|7[0-8]|8\d|9[189])\d{8}$
电子邮箱	^[a-zA-Z0-9.!#$%&'*+/=?^_`{\|}~-]+@[a-zA-Z0-9](?:[a-zA-Z0-9-]{0,61}[a-zA-Z0-9])?(?:\.[a-zA-Z0-9](?:[a-zA-Z0-9-]{0,61}[a-zA-Z0-9])?)*\.[a-zA-Z]{2,}$

2）改进型 AC 自动机算法

为提升敏感信息识别效率，将知识库的敏感数据模式串预载入改进型 AC 自动机的 Trie 树[11]。该树采用分层结构，新增模式串时仅更新对应层级节点，通过路径标记记录分支变化，利用字符级哈希表建立子节点映射，使模式匹配时间复杂度降至 O(1)。例如，当匹配身份证号前缀“320”时，传统 AC 自动机需多次回溯失败指针（沿失败指针回溯至“0”→“2”→“3”），而改进型算法可基于已匹配的前缀长度，直接跳转至对应子节点，从而减少冗余计算。

3）算法融合

为兼顾识别效率与人工维护成本，采用正则表达式与改进型 AC 自动机融合的识别机制。其总体框架如图 5 所示。针对固定格式敏感字段，先利用正则表达式快速过滤无效数据，并将初步筛选之后的数据输入改进型 AC 自动机模型，再利用哈希加速状态转移与最长前缀跳跃策略，匹配多种敏感数据模式串，从而实现结构化数据的敏感信息识别。基于铁路客运结构化数据的测试表明，该机制在千万级数据中，每万条数据的匹配耗时仅为 35 ms，其效率显著优于传统正则匹配，且支持敏感字段库的动态扩展。

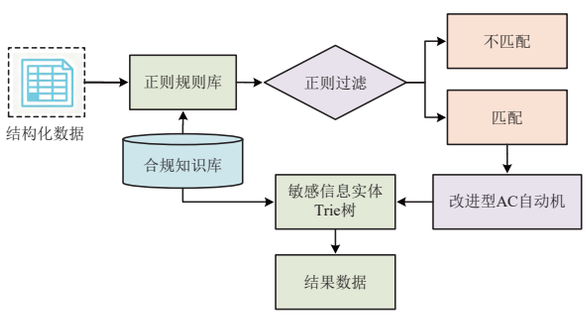

图 5 融合的识别机制框架

3.2 非结构化文本数据敏感信息识别算法设计

针对铁路非结构化文本中隐含敏感信息识别问题，设计了识别算法，该算法先将非结构化文本数据进行短文本提取，利用知识增强的 BERT 短文本敏感信息分类算法[12]，再将知识库内的客运领域知识与 BERT 融合，采用知识适配器将铁路知识库的敏感信息字典库知识集成到 BERT 底层，从而实现知识增强。该算法主要包含以下 2 个核心模块。

1）BERT 基础模型

首先，基于知识库的信息对 BERT 模型进行预训练，通过自注意力机制捕捉文本信息中各类实体的共现关系，如：使“身份证”的向量与“证件”“敏感个人信息”等词的向量在高维空间中距离更近，从而让模型学习到“身份证”属于“证件”及“敏感个人信息”的语义关联。非结构化数据敏感信息识别模型架构如图 6 所示。以一段待检测的 12306 客服对话文本为例，系统首先对其进行 3 层嵌入处理，将离散的文本 token 转化为连续的向量表示，以保留词义、位置和片段信息。其中：TEmbedding 将每个词映射为向量，捕捉通用语义；SEmbedding 用于区分不同文本片段（图 6 示例中为单段文本，故向量全部用 0 表示）；PEmbedding 用于编码词的位置信息，以建模 token 顺序关系的位置向量表，其计算式为

$$P_{p,2i}=\sin\frac{p}{10^{4(2i/d_{\text{model}})}} \tag{1}$$

$$P_{p,2i+1}=\cos\frac{p}{10^{4(2i/d_{\text{model}})}} \tag{2}$$

式中：p为token在序列中的位置索引，取值为0，1，…，$N-1$；i为维度索引；$P_{p,2i}$和$P_{p,2i+1}$分别为第p个token在偶数维度和奇数维度的位置编码值；d_{model}为隐藏维度。

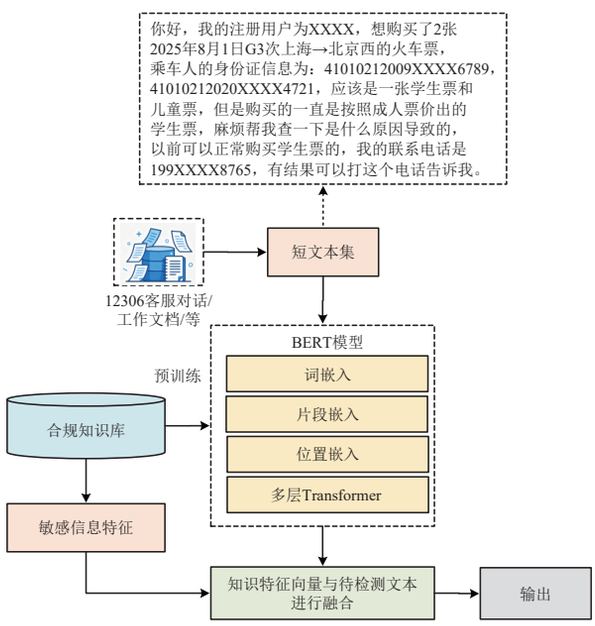

图6　非结构化数据敏感信息识别模型架构

经过3种嵌入的叠加，得到初始特征矩阵$\boldsymbol{E}$为

$$\boldsymbol{E}=\boldsymbol{T}_{\text{Embed}}+\boldsymbol{S}_{\text{Embed}}+\boldsymbol{P}_{\text{Embed}} \tag{3}$$

式中：$\boldsymbol{T}_{\text{Embed}}$为Token嵌入矩阵，用于将每Token映射为向量表征以捕捉通用语义；$\boldsymbol{S}_{\text{Embed}}$为Segment嵌入矩阵，用于区分不同文本片段；$\boldsymbol{P}_{\text{Embed}}$为位置嵌入矩阵，用于表示token在序列中的位置。

接着，通过多层Transformer中的自注意力机制在文本处理时动态计算每个词与其他词之间的关联程度，从而获取长距离依赖关系，通过Transformer的多层堆叠，逐层抽象出文本的语义特征。在每一层Transformer中，对上一层的输出数据进行编码更新，表达式为

$$\boldsymbol{H}_l=f_{\text{Transformer}}(\boldsymbol{H}_{l-1})\quad l=1,2,3,\cdots,L \tag{4}$$

式中：$\boldsymbol{H}_l$为第l层的输出数据；$\boldsymbol{H}_{l-1}$为第$l-1$层的输出数据；L为总层数。

对$\boldsymbol{H}_{l-1}$进行线性变换生成查询向量$\boldsymbol{Q}$、键向量$\boldsymbol{K}$和值向量$\boldsymbol{V}$，计算其注意力得分$\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$为

$$\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=f_{\text{Softmax}}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

式中：$\sqrt{d_k}$为缩放因子；$f_{\text{Softmax}}(\cdot)$为归一化函数。

并行执行12次上述过程，将拼接结果经线性变换后，依次通过残差连接与层归一化处理完成特征融合。其组合计算式为

$$\boldsymbol{A}_o=L(M(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})+\boldsymbol{H}_{l-1}) \tag{6}$$

其中，

$$L(x)=\alpha\odot\frac{x-\mu}{\sqrt{\delta^2+\varepsilon}}+\beta$$

式中：$M(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$为多头注意力的输出；$L(\cdot)$为层归一化操作；$\mu$和$\delta$分别为特征维的均值和标准差；$\alpha$和$\beta$分别为特征维的缩放参数和平移参数；$\varepsilon$为数据稳定项，一般为极小常数；$\odot$为逐元素乘积运算。

通过上述多层特征编码与融合，最终实现语义特征从局部关联向全局意图的识别，识别结果见表3。

表3　意图识别结果

Transformer层数	特征抽象重点
1—4	局部短语关联："G3次"与"上海→北京西"； "身份证"与两个证件号的对应关系； "学生票""儿童票""成人票价"的近邻关联。
5—8	跨句逻辑关联："购买了2张…但按成人票价出"的矛盾关系； "联系电话"与"199XXXX8765"的绑定； "以前可以正常购买"与当前异常的对比。
9—12	整体意图："咨询学生票购票异常原因，并提供身份信息和联系方式"； 核心实体：身份证号、手机号、车次、时间、上下车站等关键信息的全局定位。

将上述处理后的特征输入前馈神经网络进一步提取特征，该网络由2层线性变换与ReLU激活函数构成，对其输出结果执行残差连接与层归一化操作，最终输出为

$$\boldsymbol{H}_l=L\left(f_{\text{FFN}}(\boldsymbol{A}_o)+\boldsymbol{A}_o\right) \tag{7}$$

其中，

$$f_{\text{FFN}}(\boldsymbol{A}_o)=f_{\text{Softmax}}(0,\boldsymbol{A}_o\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2$$

式中：$f_{\text{FFN}}(\boldsymbol{A}_o)$为FFN的输出；$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别为FFN第1层和第2层线性变换的权重矩阵；$\boldsymbol{b}_1$和$\boldsymbol{b}_2$分别为FFN第1层和第2层线性变换的偏置项；$f_{\text{Softmax}}$为激活函数。

最终，取序列中第1个token对应的输出（即[CLS]）作为整个文本的特征向量，记为$\boldsymbol{H}_c^l$。

2）知识适配器

为增强模型对铁路领域敏感信息的识别能力，设计知识适配器，通过知识库特征提取与注意力机制融合，将铁路领域及合规知识注入BERT基础模

型，实现领域知识与文本语义的深度结合。首先，从数据知识库中提取敏感信息特征向量$\boldsymbol{K}_{\text{init}}\in R^{d_k}$，计算敏感词表中词的加权向量和$\boldsymbol{K}_w$，其计算式为

$$\boldsymbol{K}_w=\sum_{w\in W}\boldsymbol{e}(w)\cdot g(w) \tag{8}$$

式中：W为敏感词集合；$\boldsymbol{e}(w)$为敏感词向量；$g(w)$为词w的敏感程度。

接着，将敏感模式库、上下文规则等通过特定编码方式转化为向量$\boldsymbol{K}_r$，其与$\boldsymbol{K}_w$共同构成知识特征向量，最后获得融合敏感词、规则等的知识特征$\boldsymbol{K}'$，其表达式为

$$\boldsymbol{K}'=\lambda_1\boldsymbol{K}_{\text{init}}+\lambda_2\boldsymbol{K}_w+\lambda_3\boldsymbol{K}_r \tag{9}$$

其中，

$$\lambda_1+\lambda_2+\lambda_3=1$$

式中：λ_1，λ_2，λ_3为可学习加权系数，分别表示原始敏感特征、敏感词特征、规则模式特征的重要程度。

通过注意力机制将领域知识特征向量与待检测文本进行融合，逐位置自适应权重φ为

$$\varphi=f_{\text{Softmax}}(\boldsymbol{W}_3\tanh(\boldsymbol{W}_4\boldsymbol{Z}+\boldsymbol{W}_5\boldsymbol{K}'+\boldsymbol{b})) \tag{10}$$

式中：$\boldsymbol{Z}$为待检测文本的输出特征；$\boldsymbol{W}_3$为将隐藏层映射至φ的权重矩阵；$\boldsymbol{W}_4$和$\boldsymbol{W}_5$为将$\boldsymbol{Z}$和$\boldsymbol{K}'$映射至隐藏层的权重矩阵；$\boldsymbol{b}$为偏置项。

权重φ反映了文本各部分与知识库知识的相关程度，将相应的知识特征向量输入BERT基础模型，最终输出为

$$\boldsymbol{H}_l'=\boldsymbol{H}_l+\eta\boldsymbol{W}_k\cdot\boldsymbol{K}'+\gamma(\varphi\boldsymbol{Z}+(1-\varphi)\boldsymbol{W}_t\cdot\boldsymbol{K}') \tag{11}$$

式中：$\boldsymbol{H}_l'$为BERT基础模型的输出；η为全局知识融合强度系数；γ为全局文本特征保留系数；$\boldsymbol{W}_k$与$\boldsymbol{W}_t$为训练权重。

3.3 多媒体数据敏感信息识别算法设计

该算法主要用于从客运作业过程中产生的多媒体图像数据中提取涉及敏感信息的文本数据以及违规词汇[13−14]。所用多媒体图像数据包括旅客护照图片、车票照片、证件扫描件、文档截图、数据查询结果截图等。多媒体图像数据敏感信息识别框架如图7所示。该算法先通过对比度增强、数据增强、图像归一化等方式进行图像预处理，再经过特征提取、序列建模、文本预测和敏感信息识别4个步骤实现敏感信息识别。具体实现路径如下。

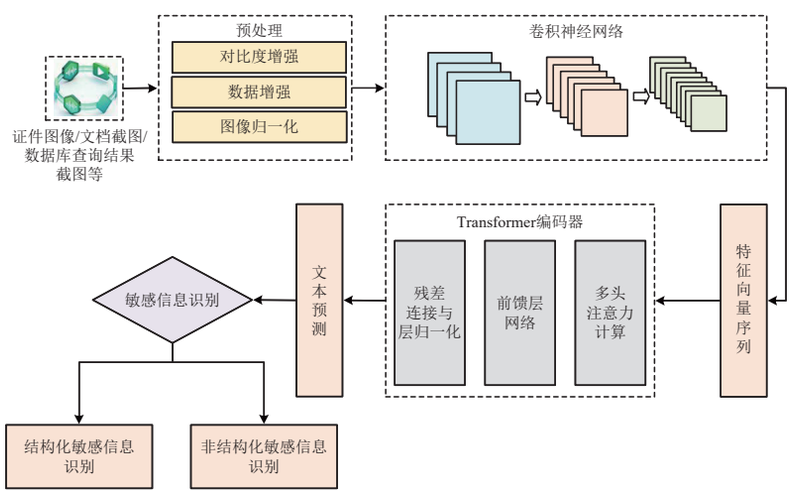

图7 多媒体图像数据敏感信息识别框架

1）特征提取

该算法以OCR-Transformer为核心实现图像文本的精准提取，利用卷积神经网络（CNN）对输入图像进行处理，提取文本笔画、轮廓等局部特征。针对多媒体图像数据中文本因拍摄角度不佳、光线条件差或文档磨损导致的模糊问题，通过CNN的多层卷积操作捕捉文本基本结构特征，并将其转化为适合后续处理的特征向量序列。主要步骤如下。

（1）输入特征$\boldsymbol{I}\in R^{H'\times W'\times C}$（$H'$为特征图高度；$W'$为特征图宽度；$C$为RGB通道数，取值为3）经CNN完成特征提取，CNN输出特征F为

$$\begin{cases}\boldsymbol{m}_i=f_{\text{RELU}}(\boldsymbol{\xi}_i\circledast\boldsymbol{m}_{i-1}+\boldsymbol{q}_i)\quad i=1,2,\cdots,n\\ F=\boldsymbol{m}_n\end{cases} \tag{12}$$

式中：$f_{\text{RELU}}(\cdot)$为逐元素非线性激活运算；$\boldsymbol{\xi}_i$为第i层卷积核参数；$\boldsymbol{q}_i$为第i层偏置项；$\circledast$为卷积操作。

（2）对输出特征$F\in R^{H''\times W''\times D}$（$H''$为特征图高度；$W''$为特征图宽度；$D$为特征通道数，取值为512）执行空间注意力加权，强化文本关键区域特征，加权后的特征F_{attn}为

$$F_{\text{attn}}=\sigma(\boldsymbol{W}_a \circledast F+\boldsymbol{b}_a)\odot F \tag{13}$$

式中：$\sigma(\cdot)$为Sigmoid函数；$\boldsymbol{W}_a$为空间注意力卷积核参数；$\boldsymbol{b}_a$为空间注意力权重矩阵的偏置项。

最终生成特征向量序列$\boldsymbol{X}=[\boldsymbol{x}_1, \cdots, \boldsymbol{x}_P]\in R^{L\times D}$（$P$为序列长度）。

2）序列建模

通过Transformer编码器及其自注意力机制，捕捉特征提取后特征向量间的长距离依赖关系，以提炼文本语义与结构信息，从而有效缓解拍摄导致的文字弯曲、倾斜及不规则字体等问题。模型可依据特征向量相关性动态分配注意力权重，从而聚焦文本关键区域。Transformer编码器采用多头自注意力机制[15]，其核心计算为

$$f_{\text{MHSA}}(\boldsymbol{X})=O_{\text{Concat}}(h_1, \cdots, h_n)\boldsymbol{\rho} \tag{14}$$

式中：$f_{\text{MHSA}}(\cdot)$为多头自注意力运算函数；h_i第i个注意力头；$O_{\text{Concat}}(\cdot)$为拼接函数，将多个注意力头进行拼接；$\boldsymbol{\rho}$为输出矩阵的参数。

通过缩放点积与Softmax函数得到全局依赖的注意力权重，经残差连接与层归一化处理后的输出为

$$\boldsymbol{X}'=L\left(\boldsymbol{X}+O_{\text{Dropout}}\left(f_{\text{MHSA}}(\boldsymbol{X})\right)\right) \tag{15}$$

式中：$\boldsymbol{X}'$为经过多头自注意力和层归一化后的特征；$O_{\text{Dropout}}(\cdot)$为随机失活函数。

输出的最终特征$\boldsymbol{X}_{\text{enc}}$为

$$\boldsymbol{X}_{\text{enc}}=L(\boldsymbol{X}'+O_{\text{Dropout}}(f_{\text{FFN}}(\boldsymbol{X}'))) \tag{16}$$

3）文本预测

解码器基于编码器输出的特征向量序列进行逐字符迭代预测生成最终文本序列，同时引入束搜索算法动态维护多条候选路径（束宽为k）并选择全局最优序列，从而提升预测结果的准确性。解码器采用自回归生成方式，计算可得第t步的概率分布$P(y_t|Y_{<t}, \boldsymbol{X}_{\text{enc}})$为

$$P(y_t|Y_{<t}, \boldsymbol{X}_{\text{enc}})=f_{\text{Softmax}}(\boldsymbol{W}_y\cdot D(Y_{<t}, \boldsymbol{X}_{\text{enc}})+\boldsymbol{b}_y) \tag{17}$$

式中：y_t为第t步待预测字符；$\boldsymbol{W}_y$为可学习的权重矩阵；$D(Y_{<t}, \boldsymbol{X}_{\text{enc}})$为解码器模块输出的特征向量；$Y_{<t}$为已生成的字符序列；$\boldsymbol{X}_{\text{enc}}$为编码器输出的整个特征向量序列；$\boldsymbol{b}_y$为可学习的偏置向量。

束搜索算法选择Top-K候选路径，最优序列Y^*的计算式为

$$Y^*=\underset{Y\in V^T}{\arg\max}\sum_{t=1}^{T}\ln P(y_t|Y_{<t}, \boldsymbol{X}_{\text{enc}}) \tag{18}$$

式中：V为字符集，T为预设的最大序列长度。

4）敏感信息检测

根据文本类型，将提取到的信息分别输入结构化与非结构化敏感信息识别模型，进行敏感信息检测。

4　基于数据血缘关系的敏感数据标签传播技术

为确保敏感数据标签传播的可审计性和准确性，需具备追溯数据流转路径的能力。当某个数据节点存在潜在的敏感数据泄露风险，或需核验其敏感数据标签的合理性时，可依托数据血缘关系追踪到数据原始来源及其全流程处理环节。通过建立详尽的数据血缘日志，记录数据在各个阶段的操作和流转情况，从而实现对数据流转路径的精确追溯，可有效防范敏感数据泄露。

4.1　数据血缘图谱构建

1）数据采集与预处理

针对铁路客运相关系统涉及的多源异构数据源，包括Sybase和Postgres等关系型数据库，HBase和Redis等非关系型数据库，HDFS分布式文件系统[16]，Gbase和SybaseIQ数据仓库以及ETL工具和Flink等数据处理组件，设计了对应的元数据采集与管理插件。铁路客运系统数据流转关系如图8所示。对采集到的数据开展清洗、去重、标准化等预处理，剔除无效、错误及重复数据，对多源异构、格式不一的数据进行统一处理，以便后续分析。

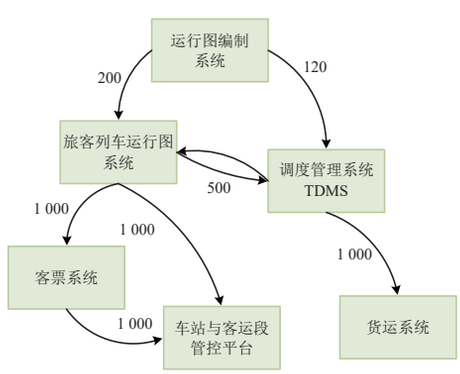

图8　铁路客运系统数据流转关系（单位：条）

2）血缘关系建模与存储

采用资源描述框架（Resource Description Framework，RDF）对数据血缘关系进行建模，以三元组（主语、谓语、宾语）的形式表示数据间的关系，进行数据血缘关系的构建[17−18]。例如，生产系统“旅客互联网订单”经ETL同步至湖仓一体

存储系统中的“旅客互联网订单表”，该表经聚合运算后，进一步生成旅客画像指标表。通过RDF三元组表示，能够清晰展示相关数据如何在不同系统、不同的存储介质间流转和转换。数据血缘关系图谱如图9所示。

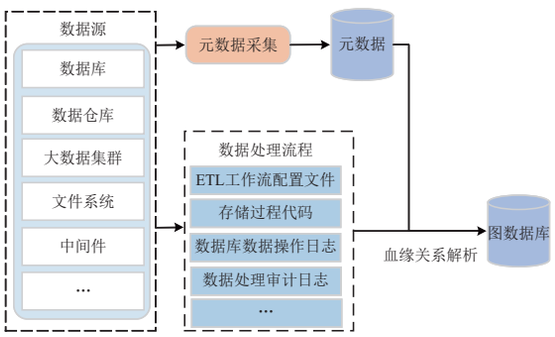

图9 数据血缘关系图谱

为保证血缘关系图谱的高效存储和查询，通过ETL任务解析、存储过程解析、SQL解析等生成三元组信息，存入开源图数据库，采用图查询语言对数据血缘关系进行查询和分析。此外，针对复杂的数据流转方式，可通过NoSQL数据库（如HBase）存储辅助性的元数据信息，进一步提升系统对复杂血缘链路的描述与检索能力。

4.2 标签传播算法设计

在铁路客运系统的数据血缘图谱上采用广度优先搜索（Breadth-First Search，BFS）算法实现敏感数据标签的传播。从标记为敏感数据的起始节点开始，逐层遍历其下游节点，并根据传播规则为每个下游节点标注敏感数据标签。例如，从包含个人身份信息的原始数据表节点出发，运用BFS算法遍历其经由ETL过程生成的所有下游数据表节点，依据传播规则为这些节点添加相应的敏感数据标签，从而确保乘客敏感信息在铁路客运全流程中得到有效保护。

为提高标签传播速度，采用并行计算技术将标签传播任务分配至多个计算节点同步执行，同时引入缓存机制存储已计算的敏感数据标签信息，避免重复计算。该算法不仅能够有效管理和保护乘客的敏感信息，还能确保整个数据生命周期内的透明度和可追溯性，从而提升敏感信息识别体系的整体安全性和可靠性。

5 试验结果及分析

为验证多模态敏感信息识别效率，设计如下3个试验场景。

场景1：抽取客票系统部分结构化数据，验证敏感信息识别效率能否支撑客票系统日均1 000万售票产生的数据量。

场景2：抽取客服对话以及业务文档，对非结构化文本信息进行验证，评估模型对语义敏感信息的识别能力。

场景3：抽取护照信息以及数据库查询结果的截图信息，验证OCR-Transformer模型的文本识别能力，并将识别出的文本分别输入对应的敏感信息识别模型进行检测，具体流程如图10所示。

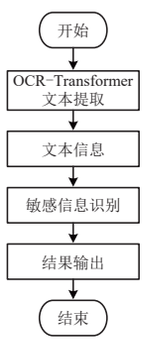

图10 场景3试验流程

5.1 环境及数据集

1）硬件环境

模型在信创环境下进行验证，采用2台服务器，每台为24核ARM架构CPU，256 G内存；500 GB SSD数据盘，4张寒武纪370显卡，所有算法均基于麒麟操作系统、Python 3.8环境运行。

2）数据集选择

场景1选取客票系统互联网售票数据，包含注册用户、常用联系人、互联网订单、电子客票等超过200个结构化字段，并从中随机抽取1.8亿条数据作为评估样本。场景2随机抽取1.8万条铁路客服对话及500份业务文档，按8∶2的比例划分为训练集与测试集；测试阶段随机选取1.44万条标注语料及400份业务文档，为保障模型训练精度，统一将非结构化文档字符集转换为UTF-8编码。场景3随机抽取护照图像5 000张、数据库查询结果截图500张，同时生成部分包含弯曲、倾斜或不规则字体的图片样本，共同用于模型的训练与验证。

5.2 结构化数据敏感信息识别算法验证

在相同的硬件环境下分别利用正则表达式匹配、传统AC自动机、改进型AC自动机以及双阶段匹配算法对200个字段的1.8亿条结构化数据进行敏感信息识别，其验证结果见表4。由表4可知，

所提双阶段识别算法处理百万级数据耗时仅4.6 s，对应检测速度可达约21.7万条·s^{-1}，约为改进型AC自动机方案的2倍，同时具有动态扩展能力，其动态规则扩展耗时缩短至127 ms，能够支持铁路业务高频变化需求，实现无中断的敏感信息规则注入。

表4　结构化数据敏感信息识别模型验证结果

方案名称	动态规则扩展耗时/ms	百万级数据处理耗时/s
正则表达式		18.0
传统AC自动机	202	10.2
改进型AC自动机	165	8.7
双阶段识别算法	127	4.6

5.3　非结构化数据敏感信息识别算法验证

针对非结构化文本，验证知识增强型BERT模型在上下文敏感场景中的识别能力。从铁路客服对话语料库中随机抽取1.5万条，标注其中的敏感实体，将知识增强型BERT模型与原始BERT-base模型、BERT＋规则后处理、行业通用BERT模型进行对比，试验结果见表5。由表5可知，知识增强型BERT模型的敏感实体F_1值达91.24%，上下文误判率降至5.88%，显著优于其他模型。

为验证知识增强型BERT模型在实际铁路客运非结构化文本中的泛化能力，构建1个包含口语化表达、同义改写、跨句上下文依赖及模板/自由文本混合等多种干扰因素的鲁棒性测试集，基于该数据集的模型F_1值对比见表6。由表6可知，与其他模型相比，知识增强型BERT模型的F_1值最高，泛化能力更强。

表5　非结构化数据敏感信息识别模型验证结果

模型名称	敏感实体F_1值/%	敏感信息上下文误判率/%
原始BERT-base	81.22	13.22
BERT＋规则后处理	84.27	10.21
通用BERT	87.92	6.73
知识增强型BERT	91.24	5.88

表6　鲁棒性测试集上模型F_1值对比　%

模型名称	F_1值
原始BERT-base	76.45
BERT＋规则后处理	80.12
通用BERT	84.05
知识增强型BERT	88.66

5.4　多媒体数据敏感信息识别模型验证

OCR-Transformer采用Encoder-Decoder架构，其中Encoder由12层ResNet和6层Transformer组成，Decoder则为12层Transformer，并在注意力机制中引入空间位置编码。训练阶段采用Adam优化器，设置其学习率为1e-4，批量大小为16，共训练20个轮次。为验证模型在复杂场景下的效果，选取包含弯曲、倾斜及不规则字体的文本进行敏感信息识别，并将OCR-Transformer与基于卷积循环的文本识别网络CRNN[19]及基于自注意力的文本识别网络SATRN进行对比。验证结果见表7。由表7可知：所提OCR-Transformer文本检测模型的字符准确率可达93.71%，模糊图像性能衰减为8.32%；在推理效率方面，OCR-Transformer推理平均耗时为128 ms，虽高于轻量级的CRNN，但相比结构更复杂的SATRN具有明显速度优势，在精度与效率之间取得了良好平衡。

表7　模型验证结果

模型名称	字符准确率/%	字段完整率/%	模糊图像性能衰减/%	推理平均耗时/ms
CRNN	80.94	62.41	−21.71	91
SATRN	87.13	64.63	−16.24	172
OCR-Transformer	93.71	91.12	−8.32	128

6　结　语

本文构建了基于知识库的多模态敏感信息协同识别体系，提出基于数据知识库的铁路客票敏感数据智能识别技术。针对结构化数据，研发了正则表达式与改进型AC自动机的双模式匹配算法，可实现约21.7万条·s^{-1}的敏感信息识别效率；针对非结构化数据，设计了知识增强型BERT模型提升语义理解能力通过领域知识图谱注入，将敏感实体识别的F_1值提升至91.24%，上下文误判率降至5.88%；针对多媒体文件，开发了OCR-Transformer的敏感信息特征提取及识别技术，图片文本提取及敏感信息识别准确率达93.71%。

通过解析数据资源元数据及数据处理流程，构建数据血缘图谱，融合图论与知识库构建技术，集成多源法规标准和行业规范，形成结构化规则引擎，从而实现数据流动合规性的智能研判。结合“识别—脱敏”联动机制与知识库驱动的语义推

理，建立覆盖数据全生命周期的合规决策支持系统，为高并发、多模态场景下的数据合规治理提供了可落地的技术方案，对数字经济时代的数据资产运营与要素市场建设具有重要的实践指导意义。

参 考 文 献

[1] 单杏花,张志强,宁斐,等.中国铁路电子客票关键技术应用与系统实现[J].中国铁道科学,2021,42(5):162-173.
(SHAN Xinghua, ZHANG Zhiqiang, NING Fei, et al. Key Technology Application and System Implementation of China Railway Electronic Ticket [J]. China Railway Science, 2021, 42 (5): 162-173. in Chinese)

[2] 符泽凡,姚竟发,滕桂法.基于BERT模型的网站敏感信息识别及其变体还原技术研究[J].现代电子技术,2024,47(23):105-112.
(FU Zefan, YAO Jingfa, TENG Guifa. Research on Website Sensitive Information Identification and Variant Restoration Technology Based on BERT Model [J]. Modern Electronics Technique, 2024, 47 (23): 105-112. in Chinese)

[3] 李扬,潘泉,杨涛.基于短文本情感分析的敏感信息识别[J].西安交通大学学报,2016,50(9):80-84.
(LI Yang, PAN Quan, YANG Tao. Sensitive Information Recognition Based on Short Text Sentiment Analysis [J]. Journal of Xi'an Jiaotong University, 2016, 50 (9): 80-84. in Chinese)

[4] LI W P, WU H Y, YANG J. Intelligent Recognition Algorithm for Social Network Sensitive Information Based on Classification Technology [J]. Discrete and Continuous Dynamical Systems-S, 2019, 12 (4/5): 1385-1398.

[5] XU G, WU X, YAO H, et al. Research on Topic Recognition of Network Sensitive Information Based on SW-LDA Model [J]. IEEE Access, 2019, 7: 21527-21538.

[6] XU G, YU Z, CHEN Z, et al. Sensitive Information Topics-Based Sentiment Analysis Method for Big Data [J]. IEEE Access, 2019, 7: 96177-96190.

[7] KRASSER S, TANG Y, GOULD J, et al. Identifying Image Spam Based on Header and File Properties Using C4.5 Decision Trees and Support Vector Machine Learning [C]// 2007 IEEE SMC Information Assurance and Security Workshop. New York: IEEE, 2007: 255-261.

[8] WANG W, WU Y, PALAIAHNAKOTE S, et al. Cloud of Line Distribution for Arbitrary Text Detection in Scene/Video/License Plate Images [C]// Advances in Multimedia Information Processing-PCM 2017. Cham: Springer International Publishing, 2018: 433-443.

[9] SHI B, BAI X, YAO C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39 (11): 2298-2304

[10] 汪庆,陈杰.深度包检测技术中的正则表达式匹配研究综述[J].网络安全技术与应用,2024(5):28-30.
(WANG Qing, CHEN Jie. A Review of Regular Expression Matching Research in Deep Packet Inspection Technology [J]. Network Security Technology & Application, 2024 (5): 28-30. in Chinese)

[11] 姜海洋,李雪菲,杨晔.基于距离比较的AC自动机并行匹配算法[J].电子与信息学报,2022,44(2):581-590.
(JIANG Haiyang, LI Xuefei, YANG Ye, et al. Distance Comparison Based Parallel Pattern Matching [J]. Journal of Electronics & Information Technology, 2022, 44 (2): 581-590. in Chinese)

[12] 杨虹,孟晓凯,俞华,等.基于BERT模型的主设备缺陷诊断方法研究[J].电力系统保护与控制,2025,53(7):155-164.
(YANG Hong, MENG Xiaokai, YU Hua, et al. Research on Primary Equipment Defect Diagnosis Method Based on the BERT Model [J]. Power System Protection and Control, 2025, 53 (7): 155-164. in Chinese)

[13] 白志程,李擎,陈鹏,等.自然场景文本检测技术研究综述[J].工程科学学报,2020,42(11):1433-1448.
(BAI Zhicheng, LI Qing, CHEN Peng, et al. Text Detection in Natural Scenes: a Literature Review [J]. Chinese Journal of Engineering, 2020, 42 (11): 1433-1448. in Chinese)

[14] KIM G, HONG T, YIM M, et al. OCR-Free Document Understanding Transformer [C]// Computer Vision-ECCV 2022. Cham: Springer Nature Switzerland, 2022: 498-517.

[15] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models from Natural Language Supervision [J]. ArXiv e-Prints, 2021: arXiv:2103.00020 [cs. CV].

[16] ISLAM N S, RAHMAN M W, JOSE J, et al. High Performance RDMA-Based Design of HDFS over InfiniBand [C]//

SC'12: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis. New York: IEEE, 2012.

[17] MONDAL S, MUKHERJEE N. Efficient NoSQL Graph Database for Storage and Access of Health Data [C]// Computer Communication, Networking and IoT. Singapore: Springer, 2021: 135-146.

[18] 潘晓华,金泳,高扬华,等.面向复杂数据审计需求的数据血缘构建方法[J].计算机应用研究,2024,41(1):76-82.
(PAN Xiaohua, JIN Yong, GAO Yanghua, et al. Data Lineage Construction Method for Complex Data Audit Requirements [J]. Application Research of Computers, 2024, 41 (1): 76-82. in Chinese)

[19] XIE Y R. Application of CRNN and OpenGL in Intelligent Landscape Design Systems Utilizing Internet of Things, Explainable Artificial Intelligence, and Drone Technology [J]. IEEE Transactions on Consumer Electronics, 2025, 71 (2): 3930-3940.

Research on the Application of Intelligent Recognition Technology for Sensitive Railway Ticket Data Based on Data Knowledge Base

HAO Xiaopei, YAN Zhiyuan, ZHANG Junfeng, LI Wen,
LIU Xiangkun, SHI Ruijun

(Institute of Computing Technology, China Academy of Railway Sciences Corporation Limited, Beijing 100081, China)

Abstract: To address the data security risks arising from the explosive growth of railway passenger transport data, the core lies in achieving intelligent identification and dynamic protection of sensitive information. Then, an intelligent identification technology for sensitive data in railway passenger tickets based on data knowledge base is proposed. Firstly, a three-level knowledge base of "laws and regulations-industry standards-enterprise norms" is constructed. Secondly, combined with historical railway passenger ticket data, a multi-level intelligent identification algorithm for sensitive data is designed, thereby efficiently and accurately identifying sensitive information in multi-modal data. On this basis, the graph technology is finally introduced to construct a data asset and sensitive data lineage graph, and based on the topological relationship of data flow, the efficient propagation of sensitive information labels among related data nodes is achieved. The results show that the sensitive information identification efficiency of the proposed technology reaches about 217 000 messages per second in structured data processing, which is almost twice as high as the traditional solution. In unstructured data processing, through domain knowledge graphs injection, the F_1 value of sensitive entity recognition is increased to 91.24%, and the context misjudgment rate is reduced to 5.88%. The accuracy of text extraction and sensitive information recognition of multimedia images reaches 93.71%. This technology can significantly improve the accuracy and processing efficiency of sensitive data identification in railway passenger tickets.

Key words: Sensitive data; Knowledge base; Railway ticket; Intelligent recognition; Label propagation; Lineage graph

(责任编辑　杨婧婕)