 DOI: https://doi.org/10.12345/bdai.v7i3.38169

Research on the Construction Method of High-Quality Police Datasets

Chunbing Chen

Zhuhai Xindehui Information Technology Co., Ltd., Guangzhou, Guangdong, 510260, China

Abstract

The construction of high-quality police datasets is still in the exploratory stage. The development path of dataset construction based on existing data governance outcomes lacks systematic design, resulting in a disconnect between datasets and actual public security application requirements. This paper proposes an engineering implementation method for constructing high-quality police datasets, covering four stages: establishment of a professional knowledge system, design of a labeling system, data annotation, and dataset partitioning and adaptation. It also innovatively introduces a police chain-of-thought annotation method to explicitly represent the reasoning path of police work, so as to improve the decision support capability of datasets for police large models.

Keywords

high-quality police datasets; artificial intelligence application; chain-of-thought annotation; knowledge system; big data

警务高质量数据集构建方法探究

陈纯冰

珠海市新德汇信息技术有限公司，中国 · 广东 广州 510260

摘　要

警务高质量数据集建设处于探索阶段，基于现有数据治理成果开展数据集建设路径缺乏体系化设计，导致数据集与公安实战需求脱节。本文设计涵盖专业知识体系搭建、标签体系设计、数据标注、数据集划分与适配四阶段的警务高质量数据集构建的工程化落地方法，并创新性引入警务思维链标注方法，以显性化警务工作推理路径，提升数据集对警务大模型决策支撑能力。

关键词

警务高质量数据集；人工智能应用；思维链标注；知识体系；大数据

1 引言

人工智能、大数据等技术的深度应用，已成为提升公安侦查效率、强化社会治安防控、保障公共安全的核心抓手，而高质量数据集的建设不仅是提升大模型性能的关键，也是推动公安“人工智能 +”行动落地的关键前提。当前，公安系统已沉淀了海量涉案数据、治安管理数据、人口基础数据等资源，并通过政务协同机制整合了教育、税务等高价值社会数据，公安大数据处理与应用已形成比较成熟的体系。然而与大数据应用成熟度鲜明对比的是，警务高质量数据集构建仍处于探索起步阶段，基于现有数据治理成果的数据集高效高质构建路径缺乏体系化设计，导致数据集与警务实战需求脱节。

本文设计“专业知识体系搭建、标签体系设计、数据标注、数据集划分与适配”的体系化落地方法并明确实施路径，为警务高质量数据集工程化实施提供可操作的实践框架。本文的研究创新点在于，在传统数据集构建流程基础上，创新性引入思维连标注方法，提升数据集对警务大模型的决策支撑能力。

【作者简介】陈纯冰（1986–），男，中国广东潮阳人，本科，从事大数据，人工智能，数据要素，智慧警务研究。

2 警务高质量数据集构建问题分析

当前警务高质量数据集建设缺乏可操作的标准化落地方法，多数建设实践处于“摸索前行”状态，主要原因为：一是现有建设多依赖单一业务部门的数据资源，未形成跨部门、全流程的建设体系，导致数据集覆盖范围有限，无法满足多场景警务实战需求；二是数据集需紧密贴合公安侦查、治安管控、指挥调度等实战场景，通用数据集的建设逻辑与警务业务需求存在偏差，无法直接适配；三是在警务领域的复杂案情推理、多线索关联分析、涉案关系研判等智能应用场景中，传统的单一标签标注模式存在明显短板，仅能标注数据表面属性，难以支撑机器对警务数据背后逻辑关系与推理过程的理解，导致无法有效支撑精准的案情推理与线索研判。

结合当前警务高质量数据集建设的研究与实践现状来看，上述问题的产生在于国内外相关研究呈现“通用领域成熟、警务领域滞后，实践探索零散、理论体系缺失”的特点。国内外学术界在通用领域及金融、医疗、工业等专业领域，已形成较为完善的高质量数据集建设理论、方法，如大数据技术标准推进委员会提出了场景驱动和数据驱动两种建设模式，以及研发管理、交付管理、运维管理和运营管理四大环节的建设路径，但针对警务领域的专项研究仍处于探索阶段，尚未形成适配警务业务特殊性的体系化研究成果。

3 警务高质量数据集四阶段构建方法

本文基于以知识为核心支撑、以思维链标注为关键方法的思路，设计警务领域专业知识体系搭建、标签体系设计与规则制定、原始数据标注、数据集划分与应用适配四阶段的警务高质量数据集构建方法。

3.1 第一阶段：专业知识体系搭建

警务领域专业知识体系具有庞杂性与多维度特征，体系的梳理与构建无法一蹴而就。为最大化并高效盘活现有警务领域结构化、非结构化数据资源，高质量数据集落地实施可优先聚焦警务专业词典、业务术语库、叙词表的构建工作，以此为基础搭建标准化的专业知识体系框架。

3.1.1 素材收集与整理

针对法律法规、警务业务操作、专业技术、档案管理等素材进行收集与整理：

法律法规类：涵盖《中华人民共和国刑法》等实体法律，《公安机关办理刑事案件程序规定》等行政规章，以及公安行业技术标准、规范等规范性文件。

警务业务操作类：包括办案指引、执法细则等业务规范手册，法律文书、案件办理卷宗等文书材料，以及接处警、案件侦查等全流程的业务规范等。

警务专业技术类：涉及法医鉴定、刑事勘查等刑事技术资料，公安大数据处理、视频图像分析等信息化技术标准。

档案与管理类：包含案件卷宗、执法台账等档案资料，公安内部人事等管理规定，以及警务培训教材等资料。

3.1.2 专业词典、业务术语库、叙词表构建

以警务业务场景为核心，系统性梳理领域专属词汇，同时制定警务专业词典的编制规范与更新机制，覆盖涉案人员、作案工具、案发现场、犯罪手法、犯罪链条等核心词汇，内容结构为“词汇 + 精准定义 + 场景示例”。

业务术语库构建聚焦各警种业务流程的标准化术语整合，如接处警流程术语、案件分类术语、侦查手段术语等，通过统一术语定义与使用规范，形成跨警种、跨业务的标准化警务业务术语体系。内容结构为“术语编号 + 定义 + 英文译名 + 所属业务板块 + 法律依据 + 关联术语”。

叙词表构建围绕警务数据检索与智能标注需求设计专用叙词表，重点建立词汇间的同义、层级、关联等语义关系，形成可被机器识别的语义网络。内容结构为“优选叙词 + 非优选叙词 + 语义关系（等同 / 层级 / 关联）+ 范畴注释”。

3.2 第二阶段：标签体系设计与规则制定

在公安大数据多年建设过程中，大多数公安机关已针对实战业务场景搭建了适配的标签体系，警务高质量数据集的建设，可在现有标签体系基础上，对标签规则的逻辑进行拓展与优化。

3.2.1 标签目录构建

标签目录构建以“人地物案组织”警务五要素为核心框架，建立基于规则的标签体系及层级化目录，同时按标签属性对各实体标签进行分类，形成标准化、可扩展的标签目录结构。

实体维度划分方面，人员实体可划分为人员、网络身份等，场所实体可划分为酒店公寓、网上场所等；物品实体可划分为车辆、前端采集设备等；案件实体可划分为刑事案件、行政案件等；组织实体可划分为企业单位、社会单位等。

标签属性分类方面，针对各实体标签采用面分类的分类方法，可按作用域（如基础、行为、关系等）、业务域（如涉黄等）、管理域（如治安等）进行多维度分类。

3.2.2 标签规则集制定

标签规则集制定可在现有标签规则逻辑基础上进行拓展，聚集标签与数据样本的精准匹配逻辑，涵盖标签与样本匹配规则、业务判定依据、术语映射标准等内容，如针对“频繁与涉黄人员同行”标签，可指定“同行人为涉黄人员 + 同行周期 + 同行次数≥ 10 次”的规则，通过量化条件实现标签的自动化判定。

3.3 第三阶段：原始数据标注

在对现有公安数据资源完成清洗、去重、脱敏、合成等预处理工作后，依据已搭建的标签体系，采用自动标注为主、人工半自动标注为辅的模式，结合智能问答、文本分类、图像识别、复杂案情推理等警务实战应用场景，对预处理后的数据开展标注工作。

标注完成后形成标准化的标签计算结果明细，即样本 / 标注数据，明细需包含包括数据标识号（唯一 ID）、关联数据标识、原始数据内容、提取特征、标签、标注方式等核心字段。

①成果物表现形式。不同应用场景标签计算结果明细表现形式如：智能问答场景形式为问答对（＜问题，答案＞或＜问题，答案，标签＞），文本 / 语音分类场景形式为单样本标注（＜特征，标签＞），图像识别场景形式为单样本标注（＜图像特征，标签＞），复杂案情推理形式为思维链标注（＜语料，标签，思维链＞）等。

②思维链标注方法。思维链（COT）标注作为一种能够显性化推理路径、还原分析逻辑的标注方法，可有效刻画警务实战业务完整思维过程。因此在数据标注阶段单独设计思维链标注方法。

维链标注的流程包括以下三大关键环节：

3.3.1 思维链构建

思维链构建通过自然语言的输出增强模型推理过程的可解释性，同时融合高能干实战中的实时动态数据和用户上下文信息，采用自动化合成和警务领域专家介入相结合方式开展：

（1）自动化合成

思维链合成，依据警务业务流程拆解具体任务，生成线性或分支型思维链，以“盗窃案嫌疑人锁定”场景为例，思维链：提取现场指纹→比对指纹库→排查嫌疑人轨迹→核查动机→锁定嫌疑人。

思维链采样，从批量合成的思维链中随机抽取样本，用于后续校验。

数据校验，依据已制定的标签规则与警务领域专业知识，对采样的思维链进行自动化逻辑校验，筛选并剔除逻辑错误的样本。

（2）警务领域专家介入

在模型微调阶段，如果通过自动化规则判断、大模型回答和人工专家判断的答案不一致时，通过流转机制由警务领域专家介入，针对复杂场景的思维链再进行定制化设计。

3.3.2 思维链改写

针对构建的初始思维链，通过自反思检测改写和多策略改写进行优化，提升思维链逻辑完整性、场景适配性和数据留存率：

（1）自反思检测改写

在 COT 合成过程中，受问题难度与模型自身能力的问题，部分思维链会出现结论与真实答案不符的情况，为避免因结果错误直接丢弃样本导致数据量不足问题，在思维链构建过后引入自反思检测改写机制：

推理解析与去重，通过 AI 解析思维链对应的结论，剔除重复的推理步骤和冗余信息，精简推理逻辑。

自反思核查，AI 依托标准推理逻辑、标签规则等警务领域专业知识，自主排查思维链的逻辑漏洞。

解析与校验，对自查后的推理链进行结构化拆解，校验各推理环节的因果关系是否成立，确保逻辑闭环。

（2）多策略改写

结合警务不同实战场景的任务需求，对思维链进行针对性改写：

结构调整，根据治安防控、刑事侦查等不同场景任务，调整思维链的线性或分支型结构。

长度控制，根据任务复杂度控制思维链步数。

事实补充，为思维链补充缺失的关键案件事实，如分析“嫌疑人轨迹”时，补充其社会关系、交通出行记录等数据。

详略调控，对思维链关键环节进行详细描述，对次要环节进行简化处理。

3.3.3 思维链校验

模型微调，根据校验结果，调整模型参数，如优化权重。

消融分析，通过控制变量法分析模型各模块对思维链生成效果影响，针对性优化模型结构。

3.4 第四阶段：数据集划分与应用适配

结合模型训练的阶段目标，遵循分层抽样、业务全覆盖、无交叉重叠的核心原则，将数据集成果进行划分：

全量专业词典 / 术语库 / 叙词表作为训练集，抽取 20% 核心术语和语义关系作为验证集，抽取 30% 术语，含同义、从属关系易混淆术语作为测试集。

全量标签目录 / 规则集作为训练集和测试集，核心规则作为验证集。

标签计算结果明细（含问答对 / 思维链样本）选取 70% 样本，按业务场景分层抽样作为训练集，15% 样本作为验证集，15% 样本（含高频易错、边缘场景等）作为测试集。

4 结语

警务高质量数据集是兼具数据准确性、业务适配性、安全合规性与知识关联性的复合型数据资源，其核心特征是能够有效提升警务领域大模型性能的数据的集合。本文提出“专业知识体系搭建、标签体系设计、数据标注、数据集划分与适配”的四阶段落地方法，并创新性引入思维链标注方法，提升数据集对警务大模型的决策支撑能力，为警务数据集工程化实施提供可操作的框架。

参考文献

[1] CCSA TC601大数据技术标准推进委员会.高质量数据集实践指南(1.0)[OL].（2025-06）[2026-01-09].获取和访问路径：大数据技术标准推进委员会微信公众号.

[2] Anderson L W, Krathwohl D R. A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives[M]. New York: Longman, 2001.

[3] CCSA TC601大数据技术标准推进委员会.面向人工智能的数据治理（DG4AI）实践指南1.0 [OL]（2024-06）[2026-01-09].获取和访问路径：大数据技术标准推进委员会微信公众号.