基于异构数据源的政法信息共享平台数据预处理分析

程　玲　聂罗娜
（江西警察学院，江西 南昌 330100）

摘　要　由于公安、检察院等司法部门之间信息共享不及时问题较为严重，极大地影响了案件侦办处理的效率，政法信息共享平台当中的数据预处理分析已然成为当务之急。此次研究的主要目的是以异构数据源为基础，提出政法信息共享平台数据预处理系统设计策略。对此，本文针对异构数据源进行了简要介绍，并结合政法信息共享平台实际功能需求，从平台开发、架构分析以及相关技术应用三个角度针对基于规则库的数据预处理系统进行分析和研究，旨在对进一步提升政法信息平台应用效果有所帮助。

关键词　异构数据源　信息共享平台　数据预处理
中图分类号：TN273　　**文献标识码**：A　　**文章编号**：1007-0745(2022)09-0025-03

大数据时代的到来，使得裁判文书、资料从传统纸质转变为了电子形式，而且信息技术的应用，也使得政法信息数据呈现出了爆发式增长态势，巨大的数据信息资源给实际工作的开展带来了较大压力。因此业内针对政法信息平台的研究主要集中在信息检索服务方面，但由于公安信息其本身的特殊性，要求其不得主动对外输出，虽然各政法部门内部的信息系统得到了集中管理，但是仍然不能够满足部门间的信息查询共享需求，为解决这一问题，文章从异构数据源角度入手，针对信息共享平台数据预处理展开分析，对于打破政法部门信息共享壁垒有着重要意义。

1 异构数据源概述

异构数据源是指不同数据库管理系统间的数据。在信息化建设的过程中，由于不同业务系统以及实际管理系统的建设时间、方式、技术水平等各不相同，而且还存在其他经济、人为等多方面因素影响，在长期积累之下，形成的大量业务数据其存储方式、管理系统等均存在较大差异，不仅存在简单的文件数据库，还存在复杂的网络数据库，这些共同形成了异构数据源。数据源的异构性主要表现在以下三个方面：第一，系统异构，即数据源所在的业务系统、数据库管理系统以及操作系统之间各不相同，而表现出的系统异构；第二，模式异构，是指数据源存储模式不同，存在关系模式、对象模式等多种形式；第三，来源异构，即数据来源不同[1]。

政法数据主要是由公安局、检察院、法院以及司法局数据共同组成。在实际进行数据信息交换的过程中，多通过接口定制开发以及人工方式进行传输共享，因此相应数据信息共享效率较低，也无法对其进行科学监控和管理，极大地增加了实际工作成本、降低了信息查询效率，对于实际工作有着不良影响。通过对政法信息的研究和调研，发现与其他行业或者部门相比，政法数据信息存在以下明显特征，使得其数据源异构性更为突出。

1. 地域性。政法数据涉及的范围相对较为广泛，而不同片区的涉案人数、案发地以及作案特点等各不相同，因此形成的数据也不同，有着极强的地域性特点。

2. 影响因素多。政法数据主要是由案件数据组成的，而案件数据会受到社会生活、季节、天气以及时间段的影响。其中以时变性较为突出，案件发生的数据特点、数据量等，与时间之间有着密切的关系，会随着时间的改变而发生变化，不仅包括每个小时、周、月，而且有着明显的季节性特征，也会随年发生改变，并伴有一定周期性，在没有受到突发事件的影响时，基本能够维持其周期性特点。

3. 数据量庞大。每年发生在全国各地的刑事案件非常多，由此产生的数据信息，包括人、时间、事件、地点以及组织、机构等，积累的数据量也非常庞大。

★基金项目：江西警察学院科研项目“政法信息共享平台的数据预处理研究”，编号：2018YB001；江西省教育厅科技项目“面向智能图像识别的深度神经网络模型的研究”，编号：GJJ202203。

4. 干扰数据多。由于数据收集的时间、方式不同，部分数据是基层人员通过人工方式获取的，如文字记录、图片拍摄等，而将数据信息录入系统的是另一部分人，因此数据录入过程中可能会存在偏差问题，影响数据的真实性，尤其是在出现突发事件时，或者关键线索无法及时获取、关联时，就会导致案件数据失去价值。

2 政法信息共享平台数据预处理

2.1 平台开发需求

基于政法数据其本身的异构特点，给政法信息共享带来了极大的影响，想要实现数据的高效共享，在进行数据信息资源整合的过程中，需要对异构数据源进行事先预处理，然后再将其引入政法信息共享平台的数据库当中，以此确保各执法部门之间能够按照实际需求以及权限等级，合理合法地获取相应政法信息，切实实现政法数据共享 [2]。

2.2 平台架构分析

政法信息共享平台数据预处理系统结构主要包括异构数据源采集以及数据预处理两个部分，政法信息共享平台搭建在信息共享区域内，信息流从公安局、法院、检察院以及司法局等各个政法部门，通过政法专线，然后穿越共享平台边界保护区，将其收集到政法信息共享平台当中，共享平台对异构数据源进行预处理，进而形成信息共享平台数据库。整个信息共享平台不仅包括元数据管理、调度管理、日志管理以及数据传输管理，同时还包括数据监控功能。

根据政法数据的异构特点，异构数据信息源的采集主要包括以下两种方式：其一为大数据量实时同步采集，其二为普通定时采集。其中，前者主要应用在数据量较大的数据源端，多用于对实时性要求较高的数据采集当中，在进行采集和抽取的过程中，需要源数据端开放高级权限；而普通定时同步采集则需要数据源端开放权限，然后定时进行高频率数据同步，若无法开放权限，则需要使用低频数据同步方式。

此外，由于政法数据来源广泛，为保障数据收集质量和效率，在进行预处理的过程中，还需另外设置规则库策略，通过对数据信息的规范化处理，以此保障数据的完整、真实和一致，为后续政法数据信息的共享奠定良好基础。

经过数据预处理后的政法数据需要存入共享平台数据库当中，为保障后续数据调取应用的便利性，数据管理的高效性，需要按照不同业务特点、要求，对数据资源库进行合理划分，以供不同业务系统使用。在进行数据使用时，需要对数据变化情况进行定时捕获、加载转换，并经过整合处理后，方可入库。

在进行数据采集、预处理、管理以及存储的过程中，系统能够自动生产相应操作日志，并通过建立监控管理平台，实现对于数据操作处理方面的管控，并对数据行为进行分析和监控预警。

基于上述方法构建的信息共享平台数据预处理系统，采用了多层可扩展框架模式，在维护管理方面有着较高的便利性，而且还具有较强的可扩展空间和能力，符合政法数据特点，以及信息共享要求。

2.3 数据预处理技术

基于政法数据其本身的异构性特点，数据预处理的主要目的就是实现数据的有机提取、整理，以及脏数据的检测和处理，以此确保被纳入数据库中的数据信息的准确性、可靠性以及完整性，为后续政法信息的共享奠定良好基础。就目前实际情况来看，数据预处理主要是借助规则函数实现的，但是此类处理工具存在可扩展性较差、动态数据预处理能力较差等方面的问题，会对数据预处理的质量和效率造成极大影响。对此，结合政法异构数据源实际情况，着重从数据预处理框架、数据抽取、整理以及数据库的设计四个方面展开分析 [3]。

2.3.1 处理框架

异构数据源下的数据预处理存在较大难度，为保障数据处理效果，提出了基于规则库的多级交互式数据预处理模式。该框架模式下的数据预处理流程主要包括以下几个步骤：第一，根据不同特定业务数据，组织行业专家以及操作人员展开访谈，并结合实际业务情况，明确第一级预处理指标，然后对错误分类信息进行整理，进而形成错误分类字典，确定预处理规则，并制定基础规则库；第二，选取相应样本数据，按照基于规则库进行二级预处理，先对样本数据集进行数据检测，并针对相应算法以及规则进行评估，从中选择最佳预处理规则，并通过数据学习、规则学习，形成动态预处理规则，以此进行数据的二级预处理；第三，三级预处理，主要是根据相应业务需求，在数据库中进行数据抽取，并结合实际抽取问题，进行算法调整、规则维护等，最后评估预处理效果，找到规则当中的漏洞，结合实际需求，在相应预处理环节当中，加入其他算法或者预处理规则等，完成预处理。

2.3.2 数据抽取

相应数据预处理规则，是在连续样本训练的基础

上建立起来的，能够有效提高后续数据抽取的质量。在进行数据抽取的过程中，通过预处理规则库进行预处理策略匹配，然后将数据分布嵌入相应的应用系统当中，除了需要对少量错误数据进行汇总处理外，大体上能够实现对于政法异构数据源的规范处理，为后续数据的进一步应用奠定了良好的基础。

在进行数据抽取时，需要基于触发词算法对文书段落进行划分，触发词主要包括开始、结束两种，在进行数据抽取的过程中，若匹配到某段落当中的开始触发词，则认为该段落开始，直至匹配到结束触发词，或者下一个开始触发词为止。然后进行关键词的抽取，抽取流程主要包括以下四个步骤：第一，对文书进行拆分，将其划分为数字、字母以及字符等不同类型；第二，在拆分后的文本当中，匹配所需要抽取的字符串，统计该字符串出现的次数，以及文书中词汇的总数量；第三，计算互信息；第四，获取候选词，进行拆分匹配后，当相邻字之间的互信息大于阈值时，继续匹配，并计算互信息值，直至匹配到的互信息值小于阈值，并将这两个字之间的字符串作为候选词；第五，计算邻接熵，通过判断邻接熵与阈值的大小关系，确定是否将其加入词表当中。

2.3.3 数据整理

在数据资源采集预处理完成之后需要将其统一收录在共享平台数据库当中，并对其进行数据信息整理，为数据的储存管理以及提取应用奠定良好基础。对此，应结合实际数据信息情况特点，构建数据标准系统，充分结合国家标准要求、部门标准要求以及省级标准要求，将现有的数据表结构、代码表、格式标准等纳入数据资源库当中。

数据结构标准方面，需要将当前政法部分的信息化标准数据结构进行全面收集，不仅包括字段中英命名、数据类型、数据长度，还应包括相应约束条件等，全部收录导入共享平台当中。在数据代码标准方面，政法系统当中的各个部门已经建立了业务系统，而且不同系统有着独属于自己的系统代码，对此，需要对现有代码表进行分析，并根据相关标准以及政法数据中心资源库，以及不同业务部门特色，制定新的代码标准，建立统一代码库管理平台。在数据格式标准方面，由于政法数据格式类型相对较多，需要针对文件、数据库等不同格式类型进行标准制定，并明确加密存储要求，如日期、时间、数据等方面的格式。此外，还需要根据国标、部标等相关标准要求，明确数据展示标准，尤其是特殊字段类型的展示，应进行统一规定管理。最后，还需要对数据标准进行定期维护管理，定期按照国标、部标等相关标准对各类数据结构、代码等在系统平台当中的标准规范进行维护，并对数据结构、代码的更新情况等进行定时监控。

2.3.4 数据库设计

数据库设计主要包括以下几个方面：第一，资源目录与任务调度控制部分表的设计，主要包括资源目录共享服务信息表，关联调度控制任务表，以及属性表、权限表和日志表等。第二，用户系统与安全审计部分表设计，主要包括用户信息表、关联日志表、权限表、安全审计表以及支持用户管理和安全审计业务方面的表。第三，点对点交换与交换调度控制部分表设计，主要包括部门资源目录表、关联调度控制数据同步表、点对点统计表，以及日志表、监控表等。第四，共享信息目录部分表设计，主要包括数据共享信息表、共享数据来源表、权限表以及记录表等。第五，接口与应用配置部分表设计，可通过分层设计方式，主要包括接口配置表、业务数据表以及查询字段表等[4]。

综上所述，政法数据信息其本身有着极强的多源异构数据特点，不仅数据来源不同，而且受到的影响因素较多，数据信息共享难度较大。因此，需要针对异构数据源，对数据预处理系统进行设计研究，基于规则库的多元数据预处理系统设计方法，能够在数据样本训练不断增加的情况下，逐渐完善规则库，提升数据预处理效果，保障数据抽取质量，而且预处理速度相对较为稳定，不会造成较大延迟影响。相信随着度异构数据源的深入研究，以及数据预处理系统的不断优化，政法信息共享平台的应用质量和效率都将会得到极大提升。

参考文献：

[1] 钱源，施佺．基于多源异构数据源的高校决策支持服务平台研究 [J]. 中国教育信息化，2020(05):50-53.
[2] 刘蓓，禄凯，程浩，等．基于异构数据融合的政务网络安全监测平台设计与实现 [J]. 信息安全研究，2020,06(06):491-498.
[3] 林瑀，陈日成，金涛．面向复杂信息系统的多源异构数据融合技术 [J]. 中国测试，2020,46(07):1-7,23.
[4] 乔伟，靳德武，王皓，等．基于云服务的煤矿水害监测大数据智能预警平台构建 [J]. 煤炭学报，2020,45(07):2619-2627.