

多源异构数据融合关键技术与政务大数据治理体系

闫佳和, 李红辉, 马英, 刘真, 张大林, 江周娴, 段宇航

引用本文

闫佳和, 李红辉, 马英, 刘真, 张大林, 江周娴, 段宇航. [多源异构数据融合关键技术与政务大数据治理体系](#)[J]. 计算机科学, 2024, 51(2): 1-14.

YAN Jiahe, LI Honghui, MA Ying, LIU Zhen, ZHANG Dalin, JIANG Zhouxian, DUAN Yuhang. [Multi-source Heterogeneous Data Fusion Technologies and Government Big Data Governance System](#) [J]. Computer Science, 2024, 51(2): 1-14.

相似文献推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[基于抗退化混沌系统和初等元胞自动机的动态S盒设计](#)

Design of Dynamic S-box Based on Anti-degradation Chaotic System and Elementary Cellular Automata

计算机科学, 2023, 50(11): 333-339. <https://doi.org/10.11896/jsjcx.220900026>

[基于超图正则化的多模态信息融合算法](#)

Multimodal Data Fusion Algorithm Based on Hypergraph Regularization

计算机科学, 2023, 50(6): 167-174. <https://doi.org/10.11896/jsjcx.220900144>

[基于BASFPA-BP的可靠性预测模型研究](#)

Study on Reliability Prediction Model Based on BASFPA-BP

计算机科学, 2023, 50(5): 31-37. <https://doi.org/10.11896/jsjcx.220900283>

[基于数据融合的商务智能与分析架构研究](#)

Framework of Business Intelligence and Analysis Based on Data Fusion

计算机科学, 2022, 49(12): 185-194. <https://doi.org/10.11896/jsjcx.211100080>

[AutoUnit:基于主动学习和预测引导的测试自动生成](#)

AutoUnit:Automatic Test Generation Based on Active Learning and Prediction Guidance

计算机科学, 2022, 49(11): 39-48. <https://doi.org/10.11896/jsjcx.220200086>

多源异构数据融合关键技术与政务大数据治理体系

闫佳和¹ 李红辉¹ 马英² 刘真¹ 张大林³ 江周娴¹ 段宇航¹

1 北京交通大学计算机与信息技术学院 北京 100044

2 国家信息中心 北京 100045

3 北京交通大学软件学院 北京 100044

(20112038@bjtu.edu.cn)

摘要 随着信息技术的飞速发展,各级政府和大型企业掌握的数据量正在以指数级别增长。然而,数据来源多样会导致格式差异,数据质量参差不齐会影响应用效果,数据分散管理会弱化关联汇集,数据形态异构会造成语义鸿沟。在此背景下,多源异构数据融合负责将来源不同的多模态数据进行有效整合,完成数据互补与关联,进而实现信息增强。目前,大多数已有研究的关注重点集中在大数据治理流程与多模态深度学习,很少有工作研究讨论完整的多源异构数据融合技术框架。因此,在综述关键技术的基础上,文中提出了一整套涵盖“数据引接-数据清洗-数据集成-数据融合”全过程的多源异构数据融合关键技术框架,并对各个环节需要解决的问题与重点任务进行介绍。然后,通过一个政务应用实例场景,给出了政务大数据治理体系的设计,以解决政务数据来源广泛、质量参差不齐、管理分散、形态异构的问题,并进一步阐述了多源异构数据融合的重要价值。最后总结全文并展望未来。

关键词: 多源异构数据;多模态数据融合;数据治理技术;政务大数据;大数据治理流程

中图分类号 TP311

Multi-source Heterogeneous Data Fusion Technologies and Government Big Data Governance System

YAN Jiahe¹, LI Honghui¹, MA Ying², LIU Zhen¹, ZHANG Dalin³, JIANG Zhouxian¹ and DUAN Yuhang¹

1 School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China

2 National Information Center, Beijing 100045, China

3 School of Software, Beijing Jiaotong University, Beijing 100044, China

Abstract With the rapid development of information technology, the data held by governments and enterprises are growing exponentially. However, the multi-source of data will lead to different formats, the low quality of data will affect the application results, the decentralized management of data will weaken integration services, and the heterogeneous modal of data will cause semantic gaps. Under this background, multi-source heterogeneous data fusion is responsible for effectively integrating multi-modal data from different sources, and then achieve information complementarity and data association, thus realizing information enhancement. At present, most studies focus on big data governance process and multi-modal deep learning, there are few works discuss integral multi-source heterogeneous data fusion framework. Therefore, based on reviewing the key technologies, this paper proposes the key technologies framework of multi-source heterogeneous data fusion that covering the processes of “data collection-data cleaning-data integration-data fusion”, and introduces the problems and tasks of each stage. Then, through an example of the government affairs application, the data governance system for government data is designed, which further explains the significance of multi-source heterogeneous data fusion. In the end, this paper is summarized and future work is prospected.

Keywords Multi-source heterogeneous data, Multi-modal data fusion, Data governance technology, Big data of government affairs, Big data governance process

1 引言

随着大数据与人工智能时代的到来,数据已经成为一种基础性战略资源,并对社会生活、经济运行和政府治理产生

深远影响。在此背景下,各级政府和大型企业纷纷推进大数据平台建设,旨在通过数据挖掘来释放数据潜在价值。然而,大数据在为人们带来便利的同时,也存在一些亟待解决的问题与挑战^[1]。例如,数据来源多样、质量参差不齐、管理分散、

到稿日期:2022-12-11 返修日期:2023-03-13

基金项目:国家重点研发计划(2019YFB2102500)

This work was supported by the National Key Research and Development Program of China(2019YFB2102500).

通信作者:李红辉(hhli@bjtu.edu.cn)

形式异构。这些问题严重制约大数据的开发和挖掘,并影响着大数据的潜在价值体现^[2-7]。

在巨大的数据资源中,政府是绝大部分数据的管理者,这意味着需要更强大的数据治理能力。2015年,国务院发布了《促进大数据发展行动纲要》,并将政务大数据治理放在主要任务的第一位。然而,政务领域也面临制约大数据发展的普遍难题,具体如下。

1)数据来源广泛。政务大数据涉及社会生活的方方面面,除了各级政府部门,还有水、电、气等公共企事业单位。各单位记录数据的格式多种多样,导致后续处理出现困难。

2)数据质量参差不齐。由于各级政务信息系统在建设和运行中对数据质量重视程度不高,使得政务数据普遍存在质量问题。常见问题包括数据错误、数据缺失、数据重复等。

3)数据管理分散。目前,政务部门之间的“数据孤岛”仍然存在,即使一部分数据实现整合,也存在形式整合、部分整合、片面整合的情况,甚至在整合之后,出现失真、失准、失效问题。

4)数据形态异构。政务部门的社会治理系统每天会产生大量的文字、图像、视频、语音等数据。这些数据蕴含着巨大价值,这将会成为数据融合和智慧决策的基础资源。

综上所述,大数据平台建设固然是时代所趋,但是,若不考虑研究多源异构数据融合的关键技术与治理体系,只是

盲目进行投入,最终会导致平台达不到预期目标。由于特殊性质,政务大数据治理不仅依赖技术,还涉及行政、管理、制度、立法等多个方面。但归根结底技术才是研究的重点。目前大多数政务部门没有成熟的数据治理工具,并且对一些技术难点认知不足,也没有明确的参照体系,只能在探索和尝试中前进。因此,本文充分利用信息技术优势,力求利用热门信息技术解决政务业务难题,为政务大数据治理提供理论依据。由于政务数据的业务属性特点明显,一些特有业务模式无法应用到其他部门。因此,本文采用技术与应用分离的思路,先提出多源异构数据融合关键技术框架,在此基础上,通过一个政务应用实例场景,设计了政务大数据治理体系。

本文立足于多源异构大数据处理关键技术,对数据引接、数据清洗、数据集成、数据融合等关键技术进行了研究综述,提出了一整套包含“数据引接-数据清洗-数据集成-数据融合”全过程的多源异构数据融合关键技术框架。

该框架具体如图1所示,数据引接阶段重点考虑数据来源多样问题,数据清洗阶段重点考虑数据质量参差问题,数据集成阶段重点考虑数据分散状态汇集问题,数据融合阶段重点考虑异构数据语义互补问题。在此基础上,通过一个政务应用实例,设计政务大数据治理体系,旨在为当下盛行的“互联网+政务大数据”模式提供一定理论思考。

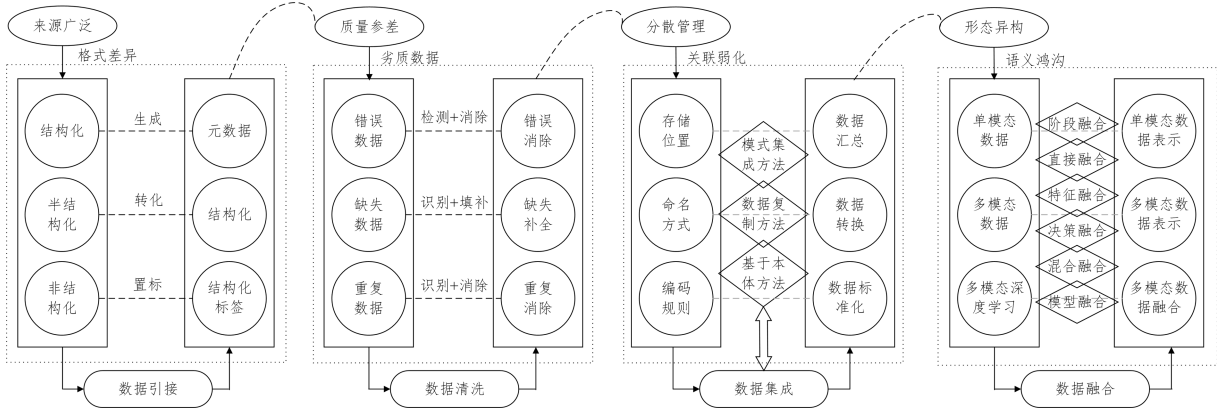


图1 多源异构数据融合关键技术框架

Fig. 1 Key technical framework of multi-source heterogeneous data fusion

本文第2章介绍了多源异构数据的相关概念和研究进展;第3章介绍了数据引接的技术要点与发展趋势;第4章介绍了数据清洗的应用背景与实现方法;第5章对数据集成的基本概念和实现模式进行了阐述;第6章介绍了多模态数据的预处理以及融合方法;第7章通过政务应用实例场景,设计了一个政务大数据治理体系;最后总结全文。

2 相关概念与研究进展

多源异构数据,是从数据来源和数据类型两个角度来衡量数据。其中,多源的含义是数据并非来自同一个信息源,而是有多个不同来源。一般在实际应用中,取自不同端口的数据集就可以被称为多源数据。异构指数据形态或格式不同,在数据领域,多模态用于表征不同形态的数据存在形式,或者同种形态的不同格式,一般包括数据库表、文本、图片、视频、音频等。在学术领域,通常认为多来源和多模态之间没有绝对关系,即多来源不等价于多模态,例如在多传感器系统中,

不同传感器产生的数据集就是多来源单模态数据集^[8]。因此,本文遵循多来源与多模态没有必然联系的原则,在数据引接阶段重点考虑数据来源不同问题,在数据融合阶段主要着力解决多模态数据异构问题。

近年来,多源异构数据融合日益成为深受研究者关注的热点问题。Li研究了多源异构数据汇集和时空融合的关键技术,提出了一套时空数据融合的技术架构,并依托于重庆市智慧城市建设平台开展诸多应用^[9]。Wei等提出了基于多尺度特征融合的多源异构遥感数据的水体提取方法,对异构遥感数据进行深层特征提取,实验表明水体提取的准确率达90.12%^[10]。Zhang等主要研究多传感器领域的多源数据融合,而且对多源数据融合模型展开综述^[11]。Feng等针对多传感器系统,利用多源异构数据融合提高传感器数据的可靠性^[12]。Mo等研究了基于联邦学习的多源异构数据融合算法,旨在解决在无交互条件下,视频、音频、文本等异构数据的融合问题。该模型包括特征提取模块、特征融合模块、特征

决策模块,最后在 MOSI 数据集上验证了模型的有效性^[13]。

由此可见,多个来源比单一来源更能提升数据可靠性,多个模态比单一模态能更全面地描述实体。然而,引入多源异构数据也带来了一些挑战,主要是各来源数据格式不同,各模态数据形态异构。虽然,目前针对多源异构数据的研究已经取得诸多进展,但是大多数工作的研究重心分别集中在两方面:结构化数据治理和多模态数据融合。很少有工作同时考虑结构化数据、半结构化数据、非结构化数据的完整处理框架。此外,政务应用场景不仅数据来源广泛、类型复杂,而且服务民生、贴近生活,非常值得进行探讨研究。

3 数据引接

数据引接指从各个业务系统或数据采集设备,将所需数据接入大数据平台的过程。数据引接是数据挖掘和应用服务的前提。在设计数据引接模块时,需要根据数据来源,考虑接入协议、格式类型、数据周期的多样性和复杂性,支持各类数据抽取传输,并提供统一的访问接口。一般常见的数据来源包括 Oracle, MySQL, DB2 等主流数据库;Webservice, Kafka, MQ 等多种类型服务接口;表格、文档、图片等各类格式文件^[14]。从宏观来看,数据类型包括结构化数据、半结构化数据和非结构化数据,如表 1 所列^[15-17]。因此,本文接下来将分别介绍结构化、半结构化、非结构化数据的引接技术要点。

表 1 结构化、半结构化、非结构化数据的特点
Table 1 Characteristics of structured, semi-structured and unstructured data

类型	结构化数据	半结构化数据	非结构化数据
概念定义	遵循某种严格的架构	具有基本的结构模式	没有固定的结构模式
一般来源	关系型数据库	各种常见类型接口	非结构数据源文件
常见实例	Oracle, MySQL, DB2	CSV, JSON, XML	文档、图片、视频、音频
使用特点	易于分析、查询、存储	支持键值调整、扩展灵活	数据丰富、体验直观
下步处理	定义元数据	转化为结构化数据	生成结构化标签

3.1 结构化数据引接

结构化数据指遵循某种严格架构的数据,有时也称作关系数据,数据库表就是最常见的结构化数据,这类数据具有易于分析、查询、存储的优点^[18]。结构化数据引接是基于构建关系数据的接入,包括 Oracle, MySQL, DB2 等关系型数据库,以及 Vertica, Greengplum 等 MPP 数据库。在结构化数据接入时,数据有全量和增量两种抽取方式,当采用增量抽取方式时,需要注意记录上一次数据抽取的最大时间,并将其作为下一次数据抽取的条件。此外,需要根据具体业务定义所抽取数据的元数据,包括列名、数据类型、顺序、长度、格式等。

3.2 半结构化数据引接

半结构化数据是介于结构化数据和非结构化数据之间的形式,其具有基本的结构模式,但不符合数据库表的形式关联结构。这类数据虽然具有相关标记用来分隔语义元素,但是数据的结构和内容混在一起,没有很好地区分开来,例如 HTML 文档、JSON、XML 等^[19]。半结构化数据引接主要是通过各类接口接入 CSV, JSON, XML, YAML 等格式的数据。在半结构化数据接入时,需要注意:配置文件属性、约定压缩格式、确定排列顺序、定义接口元数据。此外,为了能更充分和方便地利用半结构化数据,可以利用正则表达式或本体概念知识库,将半结构化数据转换为结构化数据。

3.3 非结构化数据引接

非结构化数据没有固定的结构模式,如文档、图片、视频、音频等,这类数据通常无法直接知道其内容,只能整体存储,因此检索不太方便。在实际应用中,80%的数据都是非结构化数据,这类数据中蕴含着巨大价值,需要采用新技术进行挖掘^[20]。非结构化数据引接主要是通过 FTP、本地文件、HDFS、S3 等非结构化数据源,接入各类文本、图片、视频、音频等文件。因为非结构化数据本身无法直接提供信息,所以需要采用数据置标的方法,为非结构化数据生成结构化标签,并对标签进行管理形成标签库。这个过程采用人工手动标注方法,耗时耗力,而采用智能算法标注会涉及多模态数据处理的许多相关技术,本文将在 6.3 节进行详细介绍。

数据引接流程如图 2 所示。

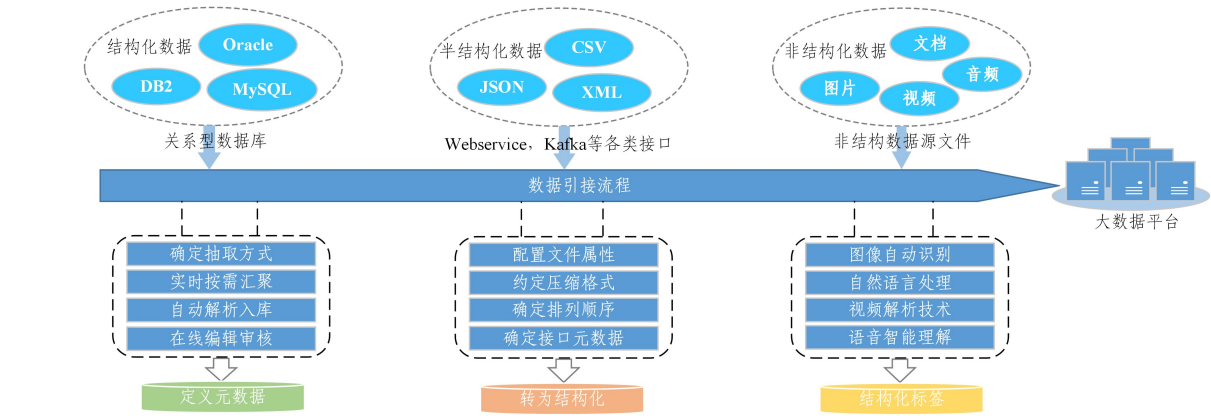


图 2 数据引接示意图

Fig. 2 Diagram of data access

通过数据引接,我们可以得到附带元数据的结构化数据、转换成结构化的半结构化数据、生成结构化标签的非结构化

数据,使得各类数据都具有结构化表征,便于存储和检索。然而,提升结构化数据质量也是不可忽视的问题。因此,第 4 章

将针对结构化数据介绍数据清洗技术。

4 数据清洗

在现实生活中,数据极易受到设备故障、传输噪声、规范差异等干扰因素的影响,从而产生一些错误数据、缺失数据、重复数据等“脏数据”。数据清洗工作就是识别出这些“脏数据”并将其去除的过程。由于数据质量的好坏直接影响后续各个数据处理过程的效果,因此数据清洗是不可缺少的步骤^[21-26]。目前,针对结构化数据的清洗技术比较成熟。从微观层面看,数据清洗技术可以分为3类:错误数据清洗、缺失数据清洗、重复记录清洗。下面将针对这3类“脏数据”,分别介绍其检测和消除技术。

4.1 错误数据清洗

4.1.1 错误数据检测

错误数据检测方法主要有两类,即基于定量检测的方法和基于定性检测的方法,如图3所示。

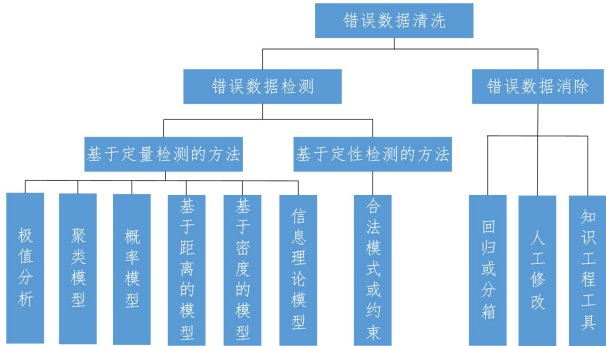


图3 错误数据清洗技术
Fig. 3 Technique of error data cleaning

基于定量检测的方法,一般是在离群点检测的基础上通过统计方法来识别出异常和误差。离群点检测是为了找出严重偏离数据分布的点。有学者又进一步将离群点检测方法分为6类:极值分析、聚类模型、概率模型、基于距离的模型、基于密度的模型、信息理论模型^[27]。

基于定性检测的方法,一般用描述性方法来指定一个合法数据的模式或约束,违反这些模式或约束的数据就是错误数据。

4.1.2 错误数据消除

错误数据消除,其目的是去除检测出来的错误数据和随机噪声。其中,随机噪声消除也叫光滑噪声技术,主要有分箱和回归等方法。错误数据消除可以借助其他材料进行人工修改,也可以使用知识工程工具找出违反规则的数据之后进行数据变化。常见的工具有 ETL 工具和数据迁移工具^[28]。

4.2 缺失数据清洗

4.2.1 缺失数据识别

在实际应用中,数据缺失是一种不可避免的现象,甚至一些系统数据的缺失率有时竟高达50%~90%,这就需要对缺失数据进行处理^[29]。如图4所示,缺失数据在数据库表中有时呈现为空白字段,有时使用特殊符号标识,例如“NULL”“#”等。然而,在一些特殊场景中,例如传感器按时间序列采样的数据,如果某个时间片的数据缺失,那么这个时间片将不会出现在数据集中。这时需要借助如 date_range() 这样

的函数功能,生成完整时间片序列,然后用完整时间片序列与原数据表做连接操作,这样就能筛选出缺失的时间片。

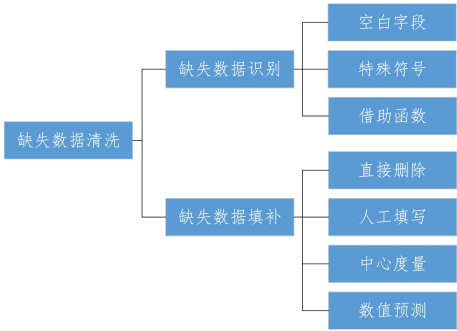


图4 缺失数据清洗技术
Fig. 4 Technique of missing data cleaning

4.2.2 缺失数据填补

识别出缺失数据后,需要对缺失数据进行填补^[30],目前有以下几种方法:直接删除、人工填写、中心度量、数值预测。直接删除方法就是直接忽略含有缺失数据的条目,当一个条目含有多个缺失数据时,常用该方法,但是直接删除会导致有价值的非缺失数据也无法发挥价值。人工填写方法最大的缺点是耗时耗力,特别是在数据集规模较大或缺失数据较多的情况下,根本无法应用。中心度量方法使用数据属性的中心度量值来填充缺失数据,例如平均值或中位数。数值预测采用常见的预测算法对缺失数据进行预测,例如回归预测、决策树、贝叶斯等方法。

4.3 重复记录清洗

4.3.1 重复记录识别

重复记录识别的关键在于,确定两条重复相似记录是否为描述同一个实体,这个过程也叫实体对齐或实体匹配。此过程不仅应用于数据清洗,也是后续数据融合的重要步骤。目前,常见的实体对齐方法有 MTransE 算法、JAPE 算法、SEEA 算法等。常用的重复性检测工具有 Febrl 系统、TAILOR 工具、WHIRL 系统、BigMatch 等,但是这些工具适合英文而不适合中文检测^[31],如图5所示。

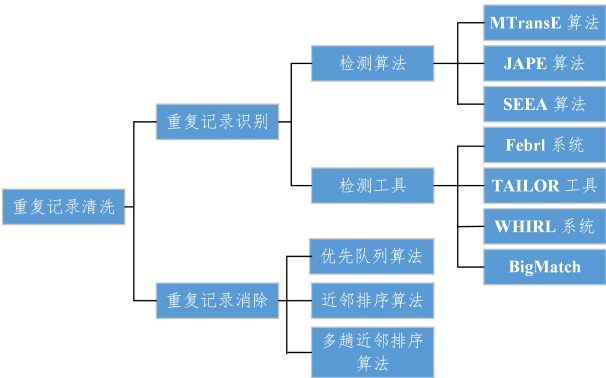


图5 重复记录清洗技术
Fig. 5 Technique of repeat record cleaning

4.3.2 重复记录消除

重复记录消除方法一般都采用先排序再合并的策略,代表算法有优先队列算法、近邻排序算法、多趟近邻排序算法等,各算法的思想详见文献^[32]。

5 数据集成

综上所述,数据引接从不同来源接入数据,数据清洗提升了数据质量。但是,此时数据还呈分散状态,主要表现为:数据存储位置不同、数据命名方式不同、数据标准规范不同。这种分散状态会导致同一实体命名不同、不同实体命名相同、数据编码不同、数据粒度不同、同业务数据未汇集等问题。因此,在数据集成阶段,主要进行数据汇总、数据转换、数据标准化等一系列工作。从宏观来看,数据集成方法主要有模式集成方法、数据复制方法和基于本体的方法^[33-41]。

5.1 模式集成方法

模式集成方法如图 6 所示,通过中介模式访问数据,为用户提供统一的查询接口。这种方法依赖于中介模式与原始模式之间的映射,相当于替用户在各个数据源进行查询。首先对各个数据源进行“包装”,即将本地查询结果转化为易于处理的表单,用户在查询时,系统会将该用户在本地的查询转化为相应数据源的查询,并将结果返回^[42]。一般来说,当数据规模较大、更新比较频繁时比较适合采用模式集成方法。

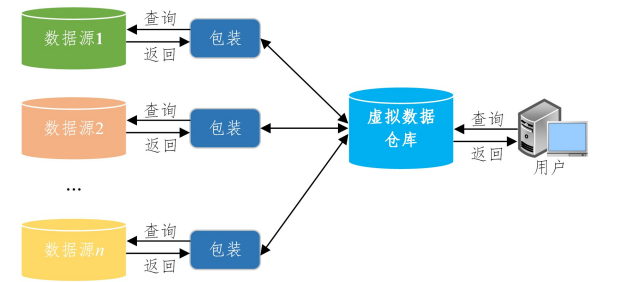


图 6 模式集成方法
Fig. 6 Pattern integration approach

5.2 数据复制方法

数据复制方法如图 7 所示,就是将用户可能用到的不同来源的数据预先复制到统一的数据源中。这种方法相当于替用户存数据,数据仓库方法就是一种常见的数据复制方法^[43]。一般来说,当数据源较稳定、用户查询有限时比较适合采用数据复制方法。

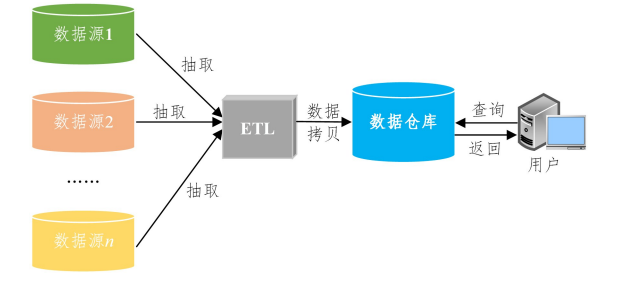


图 7 数据复制方法
Fig. 7 Data replication approach

5.3 基于本体的方法

基于本体的方法着力解决语义冲突问题。当发生语义冲突时,可以使用冲突忽略、冲突避免和冲突消解这 3 类策略,这些策略需要用到基于本体的方法。基于本体的数据集成方法可以分为单本体方法、多本体方法和混合本体方法,如表 2

所列。其中,单本体方法要求所有数据源都要与共享词汇库的全局本体相关联。多本体方法要求每个数据源都用各自的本体进行描述,而不是共享词汇库。混合方法中每个数据源用各自的本体进行描述,并且建立一个共享词汇库,能够有效解决多来源数据的语义冲突问题^[44]。

表 2 基于本体的数据集成方法的特点
Table 2 Characteristics of data integration method based on ontology

方法	单本体方法	多本体方法	混合本体方法
是否有共享词汇库	是	否	是
是否有各数据源本体	否	是	是
优点	体现全局性	可灵活调整	有效解决语义问题
缺点	应用范围小	共享一致性差	实现较为复杂

6 数据融合

数据融合与数据集成不同,数据集成表示物理位置的集中,并将这些数据统一地提供给用户。数据集成是提供数据融合所需完整数据的前提步骤。数据集成的结果是生成一个较大数据集或者有序信息序列,但是总的信息量不会增加。数据融合是通过对多个数据集进行综合处理而产生新的数据,会伴随信息的增强或损失。

6.1 数据融合流程

目前普遍认为数据融合的主要支撑技术包括模式匹配、实体对齐、冲突解决、关系推演、实体融合等^[45-48]。其中,模式匹配主要用于确定多源异构数据的属性对应关系,即查看属性名称、属性类型、属性值是否一致。实体对齐是在模式匹配的基础上,将指代同一实体的记录连接起来。冲突解决的目的是消除不同来源数据之间的同一实体的属性值冲突问题。关系推演用于确定两个实体之间的关系。实体融合是将不同来源的数据进行综合处理和多层次关联,从而提取出更加丰富、统一、精确的数据^[49-50]。数据融合流程如图 8 所示,其中非结构化数据融合涉及多模态数据融合,实现较为复杂,将在 6.3 节详细叙述。

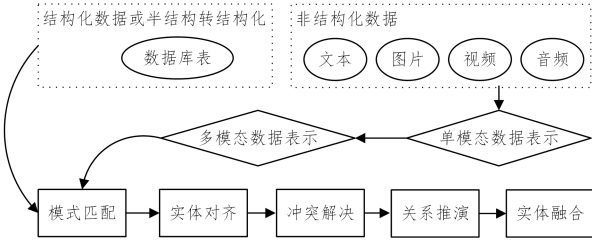


图 8 数据融合流程
Fig. 8 Process of data fusion

6.2 数据融合策略

数据融合策略可以分为阶段融合、直接融合、特征融合、决策融合、混合融合和模型融合。

6.2.1 阶段融合

阶段融合是把整体目标分解为几个独立的阶段目标,在不同阶段采用不同的数据集,使得多个数据集为同一个整体目标服务,如图 9 所示。例如,Zheng 等在城市路网设计研究

中将工作目标分为两个阶段,第一阶段通过主要交通道路将城市划分成区域,第二阶段将出租车的 GPS 轨迹映射到划分的区域上,以检测城市设计是否合理^[51]。阶段融合的优点是流程明确、便于实施,不足之处是各个阶段的数据之间没有交互。

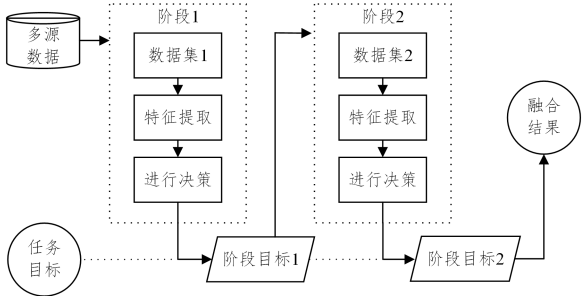


图 9 阶段融合示意图
Fig. 9 Diagram of stage fusion

6.2.2 直接融合

直接融合就是数据层融合,是将不同来源的数据直接串联拼接成一个数据集,有时也使用线性加权方法拼接^[52],如图 10 所示。这种方法是针对同一个实体进行属性补充,例如将不同传感器采集到的温度、湿度等数据直接组合为环境信息。这种方法的优点是直观简单、易于实现,缺点是没有对数据进行深入挖掘,可能引入冗余或无效信息。

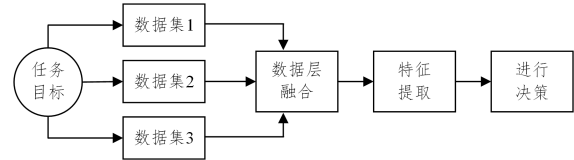


图 10 直接融合(数据层融合)示意图
Fig. 10 Diagram of direct fusion(data-level fusion)

6.2.3 特征融合

特征融合是先将各个来源的数据进行特征提取,再将这些特征分类、聚集、综合的过程,如图 11 所示。其中,特征提取应是原始信息的充分表示量或充分统计量,特征质量会影响融合结果^[53]。这种方法的优点是可以在早期利用不同来源数据的多特征相关性,不足之处是所有特征需要以相同的格式表示,容易忽视各个模态数据独有的表征、分布、密度。

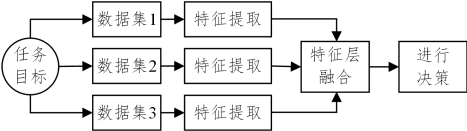


图 11 特征层融合示意图
Fig. 11 Diagram of feature-level fusion

6.2.4 决策融合

决策融合指对不同来源的数据分别使用不同模型,做出决策之后再决策结果融合,如图 12 所示。这种方法可以根据多模态特性,选择合适的方法进行建模^[54]。例如,在情感识别任务中,文本使用 LSTM(Long Short-Term Memory, LSTM)模型,音频使用隐马尔可夫模型(Hidden Markov Model, HMM),图像使用支持向量机(Support Vector Ma-

chines, SVM)。这种方法的优点是可以使用最适合的方法针对不同特性数据建模,能处理简单数据异步性。但是,数据在低水平层面的相关性被忽视,此外融合难度较高,各个模型独立训练,导致融合过程耗时耗力。

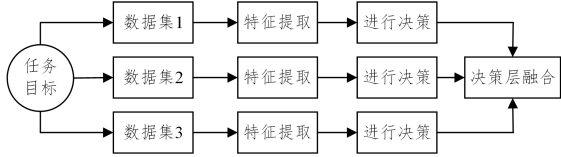


图 12 决策层融合示意图
Fig. 12 Diagram of decision-level fusion

6.2.5 混合融合

混合融合就是综合了特征融合和决策融合的数据融合方法,如图 13 所示。这种方法同时继承了特征融合和决策融合的优点,但是也增大了融合方法的复杂度。例如, Ni 等在多媒体分析任务中,采用混合融合,提出了多重 BP(Back Propagation)网络的图像识别方法,该方法综合单个 BP 神经网络的图像识别结果,提高了图像识别的性能^[55]。

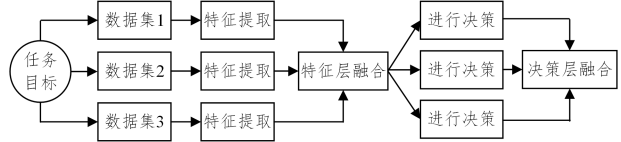


图 13 混合融合示意图
Fig. 13 Diagram of hybrid fusion

6.2.6 模型融合

模型融合,就是结合数据融合的特殊要求,对现有模型进行加工改造,融合多个模型来满足数据融合的需要,如图 14 所示。例如,通过组合 CNN(Convolutional Neural Network)和 RNN(Recurrent Neural Network)模型,可以很好地将数据的时空属性融合在一起^[56]。这种方法应用范围较广,而且效果较好,当前越来越多的研究倾向于使用此种方法。

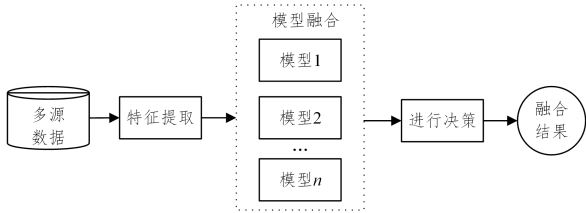


图 14 模型融合示意图
Fig. 14 Diagram of model fusion

上述融合策略各有优势和不足,具体对比如表 3 所列。

表 3 数据融合策略对比

方法	信息损失	融合难度	容错性	融合水平	融合阶段
阶段融合	小	易	差	低	各阶段执行完
直接融合	小	易	差	低	数据准备好后
特征融合	中	中	差	低	提取特征后
决策融合	大	中	中	中	做出决策后
混合融合	大	难	好	高	特征与决策时
模型融合	大	难	好	高	模型设计时

6.3 多模态数据融合

在现实生活中,人类通过视觉、听觉、味觉、触觉来感受客观世界,进而综合各种角度来描述一个实体。当前人工智能技术日新月异,人们希望计算机也能像人类那样对各种模态的数据进行综合处理,因此多模态数据融合成为当下数据融合领域的研究热点和技术难点^[57-59]。信息的不同存在形式均可以被称为一种模态,例如文本、图片、视频、音频等。虽然多模态数据在形式上表现出异构性,但是同一实体的多模态

数据在语义上则呈现一致性。通过多模态数据融合,可以实现各模态信息的互补和交流,从而更全面地展现客观实体^[60-62]。当前多模态数据融合的一般处理流程如图 15 所示。首先,根据各个模态特性选择合适的预处理方法和表征模型,即单模态数据表示;然后,在单模态数据表示的基础上,进行各个模态的综合表示,即多模态数据表示;最后,采用当前最热门的深度学习模型对这些综合特征进行融合,得到融合结果。接下来将详细介绍上述步骤。

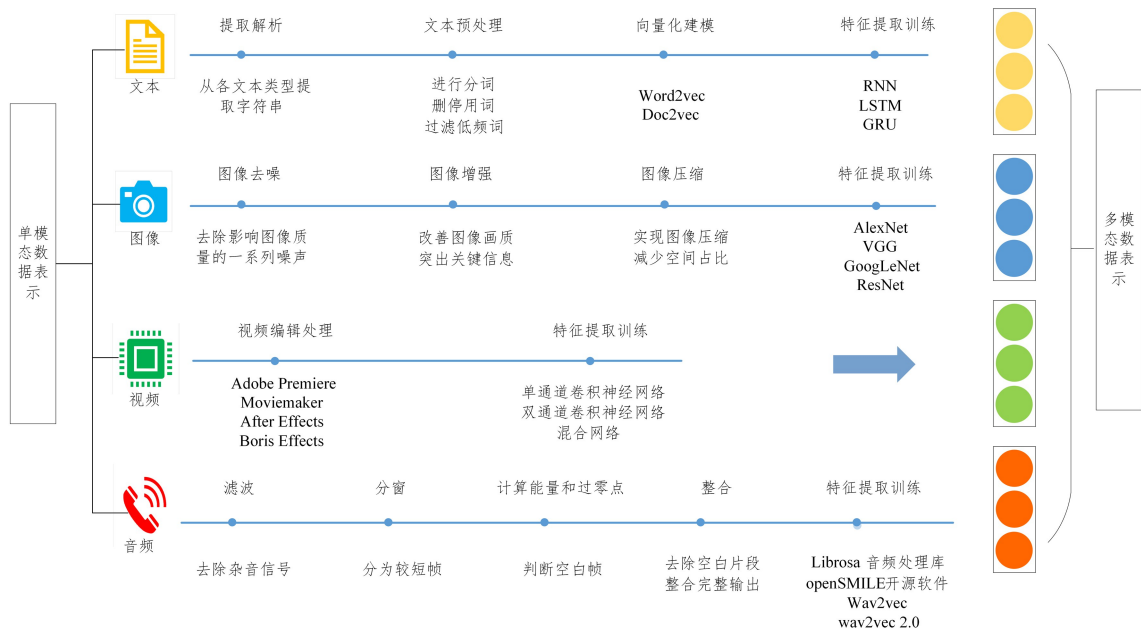


图 15 多模态数据融合示意图

Fig. 15 Diagram of multi-modal data fusion

6.3.1 单模态数据表示

在多模态数据融合领域,模态表示,即提取一个模态数据的语义特征,一直是一个充满挑战的问题。而且,不同模态数据的存在形式、处理方法和适用模型都各不相同。因此,一般先进行单模态数据表示,为多模态数据表示提供了坚实的基础。本节主要介绍文本、图片、视频、音频等模态数据。

1) 文本模态

文本是使用最广泛的数据形式,在计算机中可以存储的文本格式有 TXT, PDF, DOC, DOCX, RTF, HTML 等。首先,需要通过文本提取解析技术从文本文档中提取字符串;然后,通过文本预处理进行分词、删除停用词、过滤低频词。因为中文不像英文那样有空格作为词与词的分界,所以分词是众多自然语言处理任务 (Natural Language Processing, NLP) 的前提和基础。目前,中文分词常用方法有基于词典匹配的方法、基于统计学习的方法和基于语义理解的方法,详见文献^[63]。

要使计算机能够理解文本内容,接下来的重点是对词进行向量化建模。目前,词向量已成为主流,主要有 Word2vec, Doc2vec, ELMO, Bert 等模型。为了提升文本表示能力,研究人员一般采用预训练方式,先在大规模语料库进行预训练,获得通用的语言表示,再根据具体任务对模型进行微调^[64-65]。利用深度学习技术,可以对词向量进行特征提取,其中循环神经网络 (Recurrent Neural Network, RNN)、长短时记忆网络

(Long Short-Term Memory, LSTM)、门控循环单元 (Gated Recurrent Unit, GRU) 非常适合处理文本这样的序列数据。

2) 图片模态

图片是客观实体最直观的描述形式,常见的图片格式有 JPG, PNG, GIF, PSD, TIF, BMP 等。常见的计算机图像处理技术包括图像去噪技术、图像增强技术、图像压缩技术等。图像去噪技术旨在去除影响图像质量的一系列噪声。图像增强技术能够进一步改善原始图像的画质,还能突出有效的关键部分信息。需要注意的是,图像增强技术并不是向原始图像增加信息,而是突出强调图像的某些关键部分。图像压缩技术旨在实现图像的压缩,从而减少空间占比,提高传输的清晰度和准确度。以上技术都可以用于图像预处理,为图像特征提取打下良好的基础。

在深度学习技术发展之前,传统的图像特征提取技术包括 HOG (Histogram of Oriented Gradient, HOG), SIFT (Scale-Invariant Feature Transform, SIFT), SURF (Speed-Up Robust Features, SURF) 等。由于图像被存储为像素矩阵,这种数值化表征可以直接适用深度学习算法^[66-68]。卷积神经网络 (Convolutional Neural Network, CNN) 就是针对图像处理特别设计并发展起来的一种神经网络,例如,近几年逐步提出的 AlexNet, VGG, GoogLeNet, ResNet 等优秀的网络模型,逐渐提高了图片特征提取精度,并可以实现目标检测、物体识别、图像分割等智能化图片处理任务。

3) 视频模态

视频在实时监控方面具有不可替代的作用,在计算机中可以存储并播放的视频格式有 MP4, AVI, FLV, MOV, MPEG, RMVB 等。随着多媒体短视频的蓬勃发展,视频编辑处理的软件有 Adobe Premiere, Moviemaker, After Effects, Boris Effects 等。目前,人们希望利用人工智能算法实现视频内容理解,这有利于监控视频自动化处理。由于视频可以看作时间维度上连续的图片序列,因此要同时考虑时间维度和空间维度的属性特征。

视频的空间特征适用于卷积神经网络,可以提取相邻像素点的空间特征;时间特征适用于长短时记忆网络,对相邻帧的运动变化做特征提取。文献[47]总结出视频模态的网络结构可以分为:单通道卷积神经网络、双通道卷积神经网络、混合网络。其中,单通道卷积神经网络只用一个卷积模型完成时间和空间两方面的特征提取;双通道卷积神经网络采用两个卷积网络分别处理时间特征和空间特征;混合网络就是拼接两个不同类型的网络模型,将一个模型的输入作为另一个模型的输出,例如 CNN-LSTM 网络,这样可以发挥不同模型的优势^[69-71]。

4) 音频模态

音频可以丰富人们的视听感受,常见的音频格式有 MP3, WAV, MIDI, WMA, AAC, CD 等。在多媒体应用领域,音频编辑处理软件有 Cool Edit, Adobe Audition, Vegas, Total-Recorder。与其他信号一样,音频在进行特征提取之前也需要预处理^[72-74],用于去除杂音片段和空白片段。此过程分为 4 个步骤:第一步是滤波,即使用滤波器去除带外杂音信号;第二步是分段,即将音频信号分为较短时间段的帧,一般使用汉宁窗、海明窗、凯赛窗等窗函数进行截断;第三步是计算音频段的能量和过零点,并根据能量和过零点来判断当前片段是否为空白状态;第四步是去除空白片段并将其他片段整合为完整音频文件输出。

文献[57]指出,音频特征可以分为低级描述特征(Low Level Descriptors, LLDs)和高级统计描述特征(High Level Statistics Descriptors, HSDs)。LLDs 包括强度、响度、音高等。HSDs 是在 LLDs 基础上统计得到的,例如平均值、最大值等。文献[47]指出,可以使用音频处理技术提取一般声学特征,例如梅尔频率倒谱系数、线性预测倒谱系数等,再利用神经网络对一般声学特征进行多级非线性映射,学习包含在这些特征中的抽象及隐含信息,得到特征高阶表示。一般来说,使用 Librosa 音频处理库和 openSMILE 开源软件可以直接提取简单的音频特征。最后,将音频特征送入深度学习模型,根据任务目标进行训练,如 wav2vec, wav2vec 2.0 是目前先进的音频模型。

6.3.2 多模态数据表示

单模态数据表示为多模态数据表示提供了基础和前提。虽然多模态数据在底层形式上呈现出异构性,但是描述主体相同的多模态数据在语义上蕴含着高度一致性。多模态表示可以在保持各模态特定语义完整性的前提下,缩小各模态异质性差距,进而从不同角度更加完整地描述实体。

多模态数据表示的目的是建立一个共享空间表示或各

模态相关性约束,本质是实现不同模态之间的某种关联。文献[75]将机器学习中的多模态表示分为联合表示和协同表示。联合表示指将各个模态特征信息一起映射到一个统一向量空间中,如图 16 所示。这种方法通过将各个独立模态单独编码,从而将各个模态信息映射到共享语义子空间。因此,联合表示的优势是融合简单,而且共享语义子空间具有语义不变性,这有利于不同模态间的相互转化,缺点是早期处理各个模态比较困难。协同表示指各个模态分别映射到各自的表示空间,但要求映射后的各模态表示空间满足一定相关性要求,如图 17 所示。协同表示要求各个模态在约束条件下实现相互协同,例如语义相同或相关对象的各个模态相似距离尽量小,不同语义或无关对象的距离尽量大。协同表示的最大优势是能够在保持各模态特征的情况下实现关联,缺点是融合难度大,同时很难实现迁移学习。

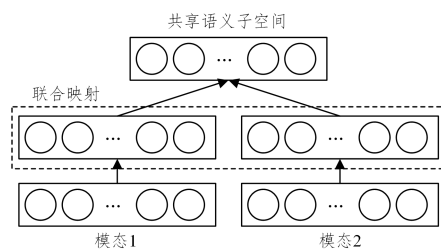


图 16 联合表示(模态共作用语义表示)

Fig. 16 Joint representation(modal coaction semantic representation)

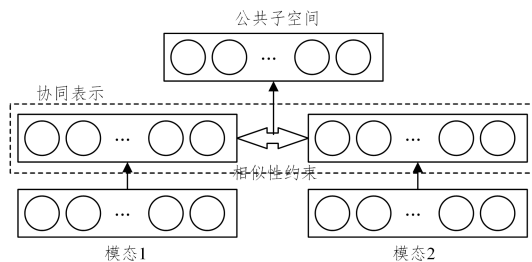


图 17 协同表示

Fig. 17 Cooperative representation

文献[47]将多模态表示分为模态共作用语义表示和模态约束语义表示。模态共作用语义表示与联合表示的定义类似,都是通过一个共享空间来融合各模态信息。模态约束语义表示与协同表示则不同,其利用一个模态的单模态表示结果来约束其他模态的表示,以使其他模态表示包含该模态的语义信息,进而实现多个模态关联,如图 18 所示。

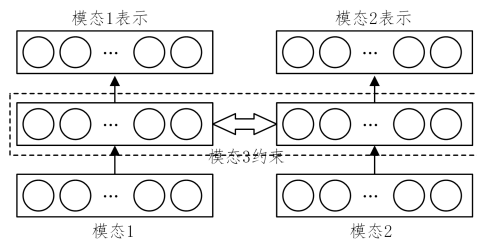


图 18 模态约束语义表示

Fig. 18 Modal constraint semantic representation

为了便于理解,文献[47]使用数学语言分别总结了模态共作用语义表示和模态约束语义表示。模态共作用语义表示

为 $\mathbf{x}_m = f(\mathbf{x}_1, \dots, \mathbf{x}_n)$, 其中, $\mathbf{x}_m$ 是模态共作用语义表示, $\mathbf{x}_1, \dots, \mathbf{x}_n$ 为各个不同模态表示, f 为神经网络模型构建的非线性映射函数。模态约束语义表示为 $f(\mathbf{W}\mathbf{x}_1)$, 其中, $\mathbf{W}$ 为映射矩阵, 负责将 $\mathbf{x}_1$ 向量映射到 $\mathbf{x}_2$ 所在的空间。

6.3.3 多模态深度学习实例

对于多模态学习而言, 需要注意不同模态的两种交互形式: 模态内关联和模态间关联。随着深度学习的发展, 多模态融合问题也有了一些较好的实现方案。Zadeh 等^[76] 所提出的记忆融合网络模型 (Memory Fusion Network, MFN), 就可以对模态内和模态间进行不同处理。如图 19 所示, MFN 由 3 部分组成, 分别是长短期记忆系统 (LSTMs)、增量记忆注意

力网络 (DMAN) 和多模态门控存储器。其中, 长短期记忆系统中每一个 LSTM 负责为一个模态信息进行编码, 其中不同模态的输入维度、输出维度、内存占用均不相同。增量记忆注意力网络的任务是负责长短期记忆系统中各个模态在时间点 t 上的跨模态交互, 其输入是一个多模态序列。多模态门控存储器由维持门 γ_1 和更新门 γ_2 两种门电路构成, γ_1 负责记录多模态门控存储器中当前的状态种类数量; γ_2 负责基于跨模态更新规则对多模态门控的内存进行更新。文献^[76] 特别指出, 由多个基准数据集上的实验对比可知, MFN 优于其他多模态学习方法, 是目前最前沿的模型之一。

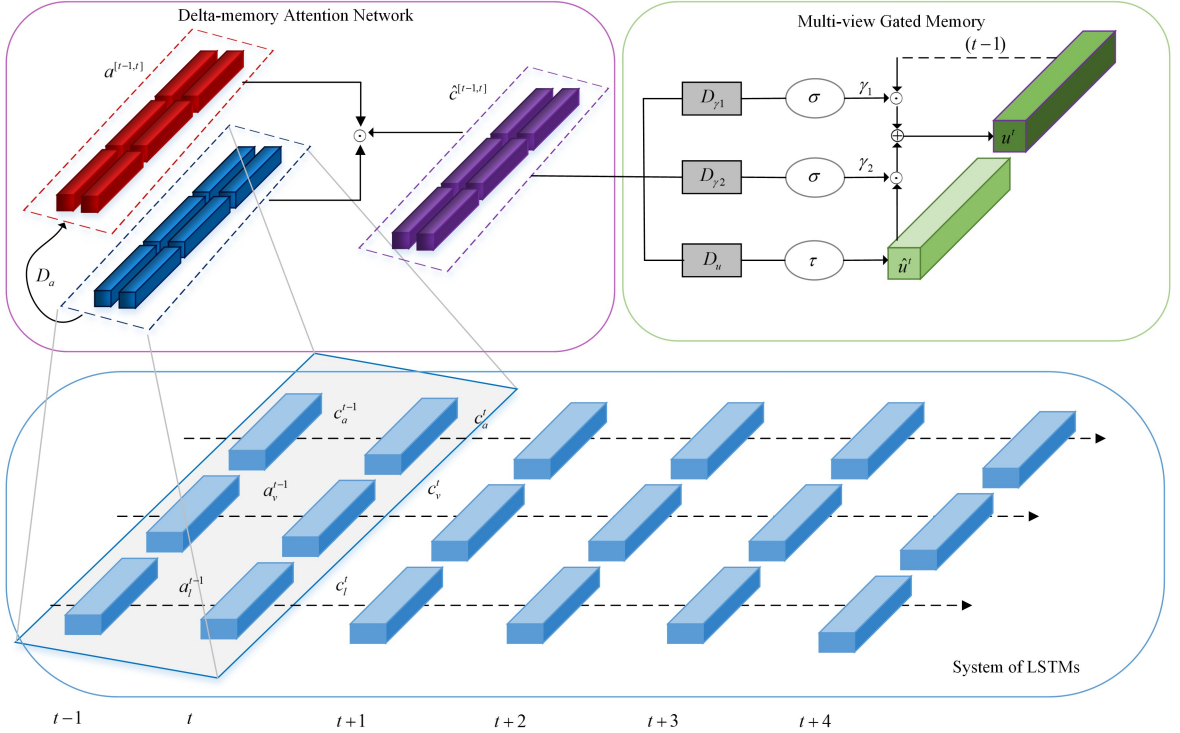


图 19 记忆融合网络模型^[76]

Fig. 19 Memory fusion network^[76]

另外, Huang 等^[77] 提出的深度结构化语义模型 (Deep Structured Semantic Model, DSSM) 使用深度神经网络 (DNN) 将两种不同模态表示为低维度的语义向量, 并 cosine 距离来计算两个语义向量间的距离, 得到语义相似度模型。如图 20 所示, 该模型由输入层、表示层和匹配层构成。输入层的任务是将各模态信息映射到一个向量空间, 并输入至 DNN 中; 表示层是利用 DNN 将高维稀疏特征映射到语义空间中的低维表示特征, 最终得到的是低维语义向量; 匹配层使用余弦相似度表示两个语义向量的相似性。

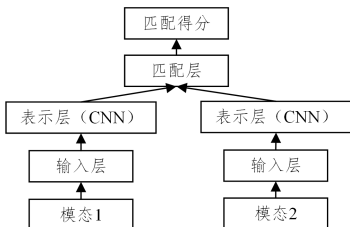


图 20 DSSM 模型的流程图^[77]

Fig. 20 Diagram of DSSM model^[77]

与上述思路不同, Wu 等^[78] 提出的多模态循环融合模型 (Multi-modal Circulant Fusion, MCF) 使用特征向量和矩阵运算来实现多模态融合, 如图 21 所示。

假设有两个不同模态的特征向量 $\mathbf{x} \in \mathbb{R}^o$, $\mathbf{y} \in \mathbb{R}^n$ 。式 (1)、式 (2) 是对向量进行投影。

$$\mathbf{V} = \mathbf{x}\mathbf{W}_1^T \quad (1)$$

$$\mathbf{C} = \mathbf{y}\mathbf{W}_2^T \quad (2)$$

其中, $\mathbf{W}_1 \in \mathbb{R}^{d \times o}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times n}$ 是投影矩阵, 负责将向量投影到低维空间。然后用得到的投影向量 $\mathbf{V} \in \mathbb{R}^d$, $\mathbf{C} \in \mathbb{R}^d$ 来构造循环矩阵 $\mathbf{A} \in \mathbb{R}^{d \times d}$ 和 $\mathbf{B} \in \mathbb{R}^{d \times d}$, 如式 (3)、式 (4) 所示:

$$\mathbf{A} = \text{cir}(\mathbf{V}) \quad (3)$$

$$\mathbf{B} = \text{cir}(\mathbf{C}) \quad (4)$$

为了使投影向量与循环矩阵的元素能够相互融合, 可以使用矩阵相乘的方法, 如式 (5)、式 (6) 所示:

$$\mathbf{F} = \mathbf{C}\mathbf{A} \quad (5)$$

$$\mathbf{G} = \mathbf{V}\mathbf{B} \quad (6)$$

也可以用循环矩阵的每个行向量分别与投影向量做哈达

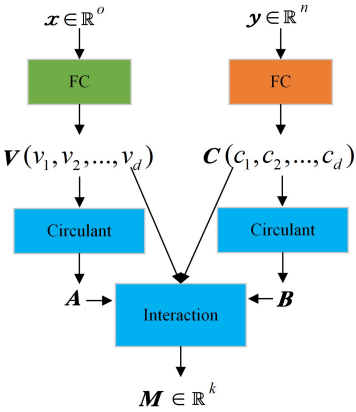
玛积,如式(7)、式(8)所示:

$$F=\frac{1}{d}\sum_{i=1}^da_i\odot C \tag{7}$$

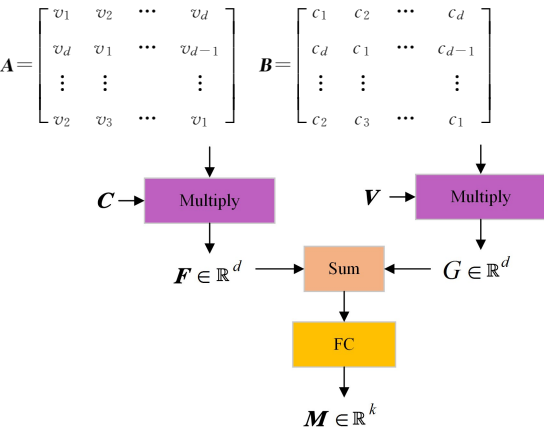
$$G=\frac{1}{d}\sum_{i=1}^db_i\odot V \tag{8}$$

其中, $a_i \in \mathbb{R}^d$ 和 $b_i \in \mathbb{R}^d$ 分别是循环矩阵 A 和 B 的行向量。

最后通过另一个投影矩阵 $W_3 \in \mathbb{R}^{d \times k}$, 将 $F \in \mathbb{R}^d$ 和 $G \in \mathbb{R}^d$ 转化为目标向量 $M \in \mathbb{R}^k$ 。



(a)多模态循环融合模型



(b)模型交互细节图

图 21 多模态循环融合模型^[78]

Fig. 21 Multi-modal cyclic fusion model^[78]

7 政务大数据治理体系

随着云计算、大数据、人工智能等信息技术不断发展,将新兴信息技术应用于政务领域,有利于提升政府的科学管理水平和社会治理能力。政务大数据记录社会生活的方方面面,其发挥价值的核心基础是充分整合利用数据资源。当前,数据治理能力不足是制约政务大数据应用建设的突出难题之一,研究探讨相关的关键技术框架和治理体系设计,对落实国家的大数据战略具有重要实践意义。本章以公安数据治理为例,根据上文综述关键技术框架,设计了一个多源异构数据融合的政务大数据治理体系,以解决政务数据来源广泛、质量参差、管理分散、形态异构的问题,充分挖掘数据的潜在价值。

政务大数据的特点是规模大、类型多、价值高,涉及公共服务领域的各类各级部门。其中,公安部门作为守护国家安全、维护社会治安、保护公民财产的重要部门,以“人、事、地、物、组织”五要素为核心,沉淀积累 PB 级的海量数据资源,

这些数据规模庞大、来源广泛、类型异构,包括结构化数据(如公安管理业务的各类数据库)、半结构化数据(如会议纪要、邮件、网络数据流)、非结构化数据(如文本、图片、视频、音频)。然而,目前这些大规模数据资源无法发挥其潜在价值。主要存在以下几个问题^[79]:1)数据治理缺乏顶层设计和总体规划,各警种部门相互独立,导致数据壁垒,进一步形成“数据孤岛”;2)原有数据处理技术落后,如缺少智能化数据采集设备、数据质量不高且难以应用,各类数据资源融合不充分;3)智能化技术应用程度不深,如大数据、人工智能等技术在公安领域的应用尚处于起步阶段。在文献查阅过程中也能发现,大多数研究集中于公安数据治理的必要性及价值性分析,对技术框架和治理体系的分析甚少。

7.1 总体治理体系

近年来,全国公安部门在基础设施建设上大力投入,包括国产自主可用 CPU 的通用服务器、网络设备、安全设备,以及各类设备运行所需要的环境(如模块化机房、云计算平台软件等)。然而,虽然公安部门已经具备强大的存储和计算能力,但是对数据价值挖掘不够重视,缺乏有效的数据治理体系。因此,本节给出数据治理体系的设计,如图 22 所示,该数据治理体系分为 4 层,分别是技术层、操作层、明细层、应用层。

1)技术层。该层包括数据存储、数据计算、数据管理的相关技术,如流式数据实时接入与分发技术、流式计算技术、离线计算技术、机器学习技术、数据抽取转换技术等。一般使用基于分布式的大数据平台,从而具有较强的存储能力和扩展能力。应用 MapReduce 批处理和 spark steaming 流式处理技术,进行任务调度、执行、管理、监控。实现数据统一管理与维护,检测数据质量,维护数据安全,监控数据血缘、提取管理元数据。基于深度学习等技术,具有自然语言处理、语音识别、图像识别、视频处理等功能,为深入挖掘公安系统中的多源异构数据打下基础。这些数据包括结构化数据(各警种业务数据库)、半结构化数据(爬虫获取的网络舆情数据流)和非结构化数据(文本记录、抓拍图片、监控视频)。

2)操作层。该层执行数据引接、数据清洗、数据集成、数据融合的关键技术。在数据引接阶段,目标是使来源广泛的数据都获得结构化表征,主要完成结构化数据的元数据提取、半结构化数据的结构化转换、非结构化数据的结构化置标。在数据清洗阶段,目标是解决数据质量参差不齐的问题,主要进行错误数据处理、缺失数据补全、重复记录消减。在数据集成阶段,目标是解决数据分散管理问题,主要包括属性集成、实体对齐、实体连接等任务。在数据融合阶段,主要解决多模态数据形态异构问题,需要包含多种融合策略,并能够处理文本、图片、视频、音频等多模态数据,能够进行单模态数据表示和多模态数据表示。使用先进的深度学习等技术,可以融合不同模态数据。

3)明细层。明细层通过操作层的相关操作,除了主题库和专题库之外,还可以得到数据规则库、数据标准库、数据标签库、数据模型库、知识图谱模型库,为应用层提供更加丰富的数据。其中,主题库和专题库是政务大数据领域的常用概念,主题库是根据某一主题需求,将基础库中的原始数据进行数据组织,即按照主题分类标准,把相关类别的数据集成在

一起。专题库是根据某一专项业务,将基础库中的原始数据按照专项业务进行组织,把某事件或活动的数据整合在一起。随着研究的深入,一些新兴的存储库被提出。例如,数据规则库存放数据处理规则,即不同部门应该按照确定的规则处理数据,避免因数据处理而引入新的不一致性。数据标准库存放数据标准体系,遵循数据元和元数据的相关标准规范,有利于数据规范建设和共享。数据标签库存放数据标签,通过数据置标的方式有利于数据快速检索和高效管理。数据

模型库存放数据模型,如概念模型和逻辑模型等。知识图谱模型库是存放数据处理加工后形成的领域知识,例如人物关系图谱。

4)应用层。应用层根据具体业务组织数据,服务于应用实际,这是数据价值最直接的体现。通常需要使用统计分析、比对碰撞、关联分析、数据挖掘等技术,这些技术可以满足公安部门常见的应用,例如线索挖掘、趋势预测、信息搜索、督察应用、布控应用、治安应用、派出所应用等。

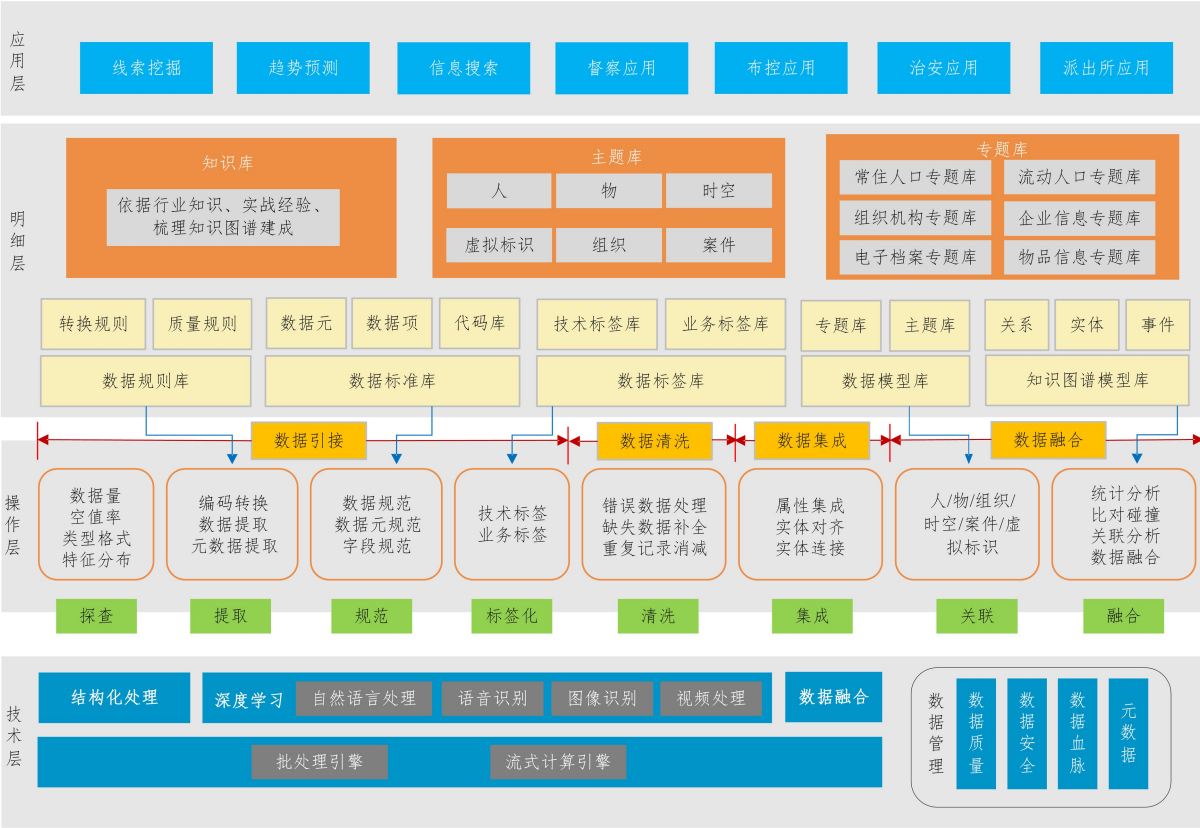


图 22 公安数据治理体系

Fig. 22 Governance system for police data

7.2 数据治理流程

接下来,结合本文关键技术框架说明具体数据处理流程^[80-81],包括数据引接、数据清洗、数据集成和数据融合。

1)数据引接。主要任务是将采集到的所有数据进行统一接入与管理,并根据数据类型完成抽取和传输,形成原始数据汇聚库。公安部门数据来源较多,包括公安业务数据、政务数据、社会数据、互联网数据及物联网数据。因此,在数据引接阶段,要根据数据的特性和用途,选择合适的方式和手段处理数据,形成标准化、易访问的原始数据汇聚库。按照本文所提关键技术框架思想,在数据引接阶段,完成结构化数据的元数据提取、半结构化数据的结构化转换、非结构化数据的结构化置标,使得来源广泛的各类数据都具有结构化表征,便于检索和访问。

2)数据清洗。主要任务是消除错误数据、补全缺失数据、去除重复数据。目前,公安业务数据很少应用数据清洗过程,“脏数据”普遍存在,要借助专业的数据清洗工具或系统,按照数据标准体系进行审查和校验,过滤掉不合规数据。按照本文所提关键技术框架思想,在数据清洗阶段,完成错误数据的

消除、缺失数据的补全、重复数据的去除。

3)数据集成。主要任务是将处于分散状态的数据进行汇集,同时完成数据汇总、数据转换、数据标准化等一系列工作。公安部门数据通常按照一个主题进行关联,如以人、物、时空、组织、案件、虚拟标识等作为主题,分别建立数据模型,能更好地服务应用。按照本文所提关键技术框架思想,在数据集成阶段可以根据主题集成相关数据。

4)数据融合。主要任务是通过多个数据集进行综合处理而产生新的数据。采用数据挖掘算法,开发数据的潜在价值,以更好地为上层应用服务。例如,将描述同一主体的数据表、文本记录、图片影像、监控视频进行信息互补与验证,还可以进一步处理形成知识图谱。

7.3 知识图谱构建

知识图谱是人工智能领域的相关概念,本质是一种语义网络,可以融合多种数据源,丰富数据语义信息,并且可以结合推理得到隐含信息,为用户提供更好的服务。

公安知识图谱将公安部门的海量多源异构数据转化为以“人、事、地、物、组织、虚拟身份”等要素为主的实体,定义并

挖掘各实体间的各种关系,有利于提前研判案情。图 23 是一个简单的知识图谱。以“人”为中心实体,建立人与“电话号码”所属关系、“人”与“护照”所属关系及“人”与“人”的关系,同时建立了“人”与“航班”的出行事件、“人”与“旅馆”的住宿事件。

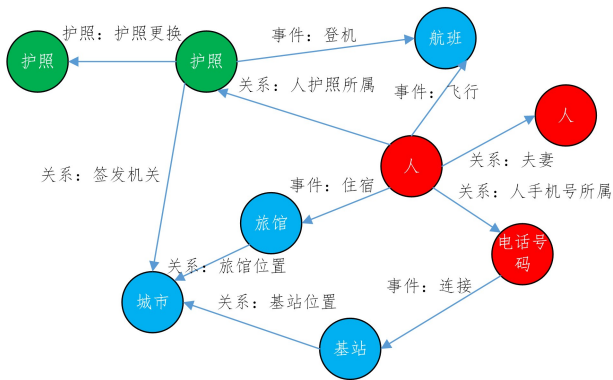


图 23 公安知识图谱示例

Fig. 23 Example of police knowledge graph

7.4 效用对比分析

本文提出的治理体系与传统治理模式的对比如表 4 所列,可以看出,本文提出的数据治理体系的理念思想更加符合当下的数据治理需求。此外,本文依托于国家重点研发计划“互联网+政务大数据透明管理与智能服务平台”,开展了多源异构数据融合相关工作,并取得了良好的效果。

表 4 本文治理体系与传统治理模式对比分析

Table 4 Comparative analysis of the proposed data governance system and traditional data governance pattern

	传统治理模式	本文治理体系
目标	将数据治理为信息	将数据治理为知识
方法	基于数据逻辑组织数据	基于机器学习挖掘价值
侧重	数据规范、标准、编目	数据清洗、集成、融合
对象	结构化数据库表	多源异构数据
结果	数据仓库	行业知识图谱

结束语 在当前时代背景下,数据来源广泛、质量参差不齐、管理分散、形态异构。因此,多源异构数据融合成为了备受关注的研究热点。本文提出了一整套包含“数据引接-数据清洗-数据集成-数据融合”全过程的多源异构数据融合关键技术框架。具体地,各环节针对的问题与重点任务如下。

在数据引接阶段,重点考虑数据来源不同问题,进行结构化数据的元数据提取、半结构化数据的结构化转换,非结构化数据的结构化置标。通过这种方式,使得来源广泛的各类数据都具有结构化表征。

在数据清洗阶段,重点考虑数据质量提升问题,使用成熟的结构化数据清洗技术处理“脏数据”,目的是消除错误数据、补全缺失数据、去除重复数据,从而有效提升数据质量。

在数据集成阶段,重点考虑将分散状态的数据进行汇总聚集,需要进行数据标准化、数据转换、数据汇总等一系列工作。主要介绍目前常见的数据集成方法,包括模式集成方法、数据复制方法和基于本体的方法。

在数据融合阶段,重点考虑数据语义信息的互补和交流。主要总结常见的数据融合策略,介绍了单模态数据表示与多模态数据表示,包括文本、图片、视频、音频等模态。最后列举

了当前热门的深度学习多模态数据融合实例。

在上述关键技术支撑下,以政务应用场景为例,提出了一个政务大数据治理体系。然而,受到问题难度和时间限制等因素制约,仍有一些问题未得到充分解决,在此特进行展望。现存问题如下:

1)数据规模控制。目前数据规模越来越大,例如仅中国某个省会城市的一小部分数据构建成数据库图,其顶点和边的个数分别达到了 10 亿和 100 亿规模,因此要根据业务重点和数据质量,合理控制数据融合的规模。

2)思路因时转变。传统数据库是先有模式后有数据,属于“自顶向下”的方式,而当前数据湖的概念是先有数据后有模式,因此需要一种“自底向上”的方式。

3)目前多模态数据融合的数据集还不是特别丰富,很少有针对同一实体的文本、图片、视频、音频等数据,也缺失相应实际场景进行有价值的研究。希望在未来开阔思路,以早日解决上述问题。

参 考 文 献

[1] MENG X F, DU Z J. Big Data Fusion Research[J]. Computer Research and Development, 2016, 53(2): 231-246.

[2] CHEN Q, ZHANG Z. Research on Data Lake Architecture of Multi-source Heterogeneous Data Governance[J]. Journal of Information, 2022, 41(5): 139-145.

[3] HUANG J F, YE P J, WANG M. Data Governance Practices Based on Big Data Basic Platform[J]. Information Technology and Standardization, 2022(6): 19-23.

[4] ZHANG A, LV N. Research on the Impact of Big Data Capabilities on Government's Smart Service Performance: Empirical Evidence From China[J]. IEEE Access, 2021, 9: 50523-50537.

[5] KHALIL M I, KIM R Y, SEO C. Challenges and Opportunities of Big Data[J]. Journal of Platform Technology, 2020, 8(2): 3-9.

[6] OTTO B. Data governance[J]. Business & Information Systems Engineering, 2011, 3(4): 241-244.

[7] WU X, ZHU X, WU G, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 97-107.

[8] YU H, LIANG Z T, YAN Y C. Research progress in multi-source and multi-modal data fusion and integration[J]. Intelligence Theory and Practice, 2020, 43(11): 169-178.

[9] LI S F. Research and application of key technologies for spatio-temporal fusion of multi-source heterogeneous data[J]. Geospatial Information, 2021, 19(10): 19-21.

[10] WEI Z H, JIA K B, JIA X W. Water extraction from multi-source heterogeneous remote sensing data based on multi-scale feature fusion[J]. Remote Sensing Information, 2021, 36(5): 41-48.

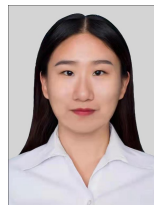
[11] ZHANG P F, LI T R, WANG G Q, et al. Multi-source Information Fusion Based on Rough Set Theory: A Review[J]. Information Fusion, 2021, 68: 85-117.

[12] FENG Y L, HU J X, DUAN R, et al. Credibility Assessment Method of Sensor Data Based on Multi-Source Heterogeneous Information Fusion[J]. Sensors, 2021, 21(7): 2542.

[13] MO H L, ZHENG H F, GAO M, et al. Multi-source heterogeneous data fusion algorithm based on federated learning[J]. Computer Research and Development, 2022, 59(2): 478-487.

- [14] BA J M, SHAO P Z, MENG Y Q, et al. Research on Big Data Governance System based on multi-mode fusion[C] // Proceedings of the 15th National Conference on Signal and Intelligent Information Processing and Application. 2022;382-389.
- [15] KUMAR K. Integrated benchmarking standard and decision support system for structured, semi structured, unstructured retail data[J]. *Wireless Networks*, 2021(Online) DOI: 10. 1007/s11276-021-02843-4.
- [16] LO G P, MUSARELLA L, SOFO G, et al. An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake[J]. *Information Sciences*, 2019, 478: 606-626.
- [17] TEKLI J. An Overview on XML Semantic Disambiguation from Unstructured Text to Semi-Structured Data: Background, Applications, and Ongoing Challenges [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2016, 28(6): 1383-1407.
- [18] DEMETROVICS J, SON HN, GUBAN A. A Formal Representation for Structured Data[J]. *Acta Polytechnica Hungarica*, 2016, 13(2): 59-76.
- [19] ZHANG L J, LI N, LI Z H. An Overview on Supervised Semi-structured Data Classification[C] // 8th IEEE International Conference on Data Science and Advanced Analytics(DSAA). 2021.
- [20] CHEN L, SHAO J, YU Z, et al. RAISE: A Whole Process Modeling Method for Unstructured Data Management[C] // IEEE First International Conference on Multimedia Big Data. 2015: 9-12.
- [21] YANG D H, LI N N, WANG H Z, et al. Optimization of Parallel Big Data Cleaning Process Based on Task Merging[J]. *Journal of Computer Science*, 2016, 39(1): 97-108.
- [22] QI Z X, WANG H Z, ZHOU X, et al. Establishment of Cost-Sensitive Decision Tree on Poor Quality Data[J]. *Journal of Software*, 2019, 30(3): 604-619.
- [23] MA Q, GU Y, ZHANG T C, et al. A Heterogeneous Multi-Source Multi-Modal Sensing Data Acquisition Method Based on Data Quality[J]. *Journal of Computer Science*, 2013, 36(10): 2120-2131.
- [24] DUAN X L, GUO B, SHEN Y, et al. Data Repair Method Based on Time Rule[J]. *Journal of Software*, 2019, 30(3): 589-603.
- [25] REDMAN T C. The impact of poor data quality on the typical enterprise[J]. *Communications of the ACM*, 1998, 41(2): 79-82.
- [26] SWARTZ N. Gartner warns firms of ‘dirty data’[J]. *Information Management*, 2007, 41(3): 6.
- [27] AGGARWAL C C. Outlier Analysis [C] // Data Mining. Cham, Springer, 2015: 237-263.
- [28] CHU X, ILYAS I F. Qualitative data cleaning . [C] // Proceedings of the VLDB Endowment. 2016; 1605-1608.
- [29] PARULIAN N N, LUDASCHER B. Towards Transparent Data Cleaning: The Data Cleaning Model Explorer (DCM/X) [C] // 21st ACM/IEEE Joint Conference on Digital Libraries(JCDL). 2021: 326-327.
- [30] HUA M, PEI J. Cleaning disguised missing data: A heuristic approach[C] // Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). New York: ACM Press, 2007: 950-958.
- [31] ELMAGARMID A K, IPEIROTIS P G, VERYKIOS V S. Duplicate record detection: A survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2007, 19(1): 1-16.
- [32] WU X D, DONG B B, DU X Z, et al. Data Governance Techniques[J]. *Journal of Software*, 2019, 30(9): 2830-2856.
- [33] VETOVA S. Big Heterogeneous Data Integration and Analysis [C] // 46th International Conference on Applications of Mathematics in Engineering and Economics(AMEE). 2021.
- [34] LI J K, WANG Y Z, LI Z A. DM_Integration: A framework for iterative large volume data integration[C] // 1st International Symposium on Data, Privacy and E-Commerce(ISDPE 2007). 2007: 68-73.
- [35] CHEN X P, YANG J, ZHANG F W, et al. Design of Data Integration Platform Based On Relational Database[C] // International Conference on Management Science and Engineering. 2010: 257-260.
- [36] MALESZKA M, NGUYEN N T. A Method for Complex Hierarchical Data Integration[J]. *Cybernetics and Systems*, 2011, 42(5): 358-378.
- [37] LIU G L, YANG C H. A Method of Data Integration Based on Cloud [C] // 2nd International Conference on Mechatronics and Control Engineering(ICMCE 2013). 2013: 1876-1879.
- [38] JARKE M, JEUSFELD M, QUIX C. Data-centric intelligent information integration-from concepts to automation[J]. *Journal of Intelligent Information Systems*, 2014, 43(3): 437-462.
- [39] GRABIS J, KAMPARS J. On-demand Data Integration for Decision-making Applications[C] // 15th International Conference on Enterprise Information Systems(ICEIS). 2013: 201-208.
- [40] ZHOU B. Data Integration as a Service for Applications Deployment on the SaaS Platform[C] // 6th International Conference on Biomedical Engineering and Informatics(BMEI). 2013: 672-676.
- [41] SAZONTEV V, STUPNIKOV S. An Extensible Approach for Materialized Big Data Integration in Distributed Computation Environments[C] // Ivannikov Memorial Workshop(IVMEM). 2019: 33-38.
- [42] HE H, MENG W, YU C, et al. Constructing interface schemas for search interfaces of web databases[C] // Proceedings of the International Conference on Web Information Systems Engineering. New York: Springer-Verlag, 2005: 29-42.
- [43] CHAUDHURI S, DAYAL U. An overview of data warehousing and OLAP technology[J]. *ACM Sigmod Record*, 1997, 26(1): 65-74.
- [44] BENEDIKT M, GRAU B C, KOSTYLEV E V. Logical foundations of information disclosure in ontology-based data integration[J]. *Artificial Intelligence*, 2018, 262(2018): 52-95.
- [45] REN Z Y, WANG Z C, KE Z W, et al. A review of multimodal data fusion[J]. *Computer Engineering and Applications*, 2021, 57(18): 49-64.
- [46] SUN Y Y, JIA Z T, ZHU H Y. A review of multimodal deep learning [J]. *Computer Engineering and Applications*, 2020, 56(21): 1-10.
- [47] LIU J W, DING X H, LUO X L. A review of multimodal deep learning[J]. *Computer application Research*, 2020, 37(6): 1601-1614.
- [48] CHEN Y, ZHOU G, LU J C. A review on the construction and application of multimodal knowledge graph[J]. *Computer application Research*, 2021, 38(12): 3535-3543.
- [49] ATREY P K, HOSSAIN M A, SADDIK A. Multimodal fusion for multimedia analysis: a survey [J]. *Multimedia Systems*, 2010, 16(6): 345-379.
- [50] YEH Y R, LIN T C, CHUNG Y Y. A novel multiple kernel learning framework for heterogeneous feature fusion and varia-

- ble selection [J]. IEEE Transactions on Multimedia, 2012, 14(3):563-574.
- [51] ZHENG Y, LIU Y, YUAN J, et al. Urban computing with taxis[C]// Proceedings of the 13th ACM International Conference on Ubiquitous Computing. 2011.
- [52] LIU W, ZHANG C, YU B. A general multi-source data fusion framework[C]// Proceedings of the 2019 11th International Conference on Machine Learning and Computing. 2019.
- [53] LIU Z K, LIU H P, HUANG W M. Visual and auditory cross-modal surface material retrieval[J]. Journal of Intelligent Systems, 2019, 14(3):423-429.
- [54] YU Y, LIN H, MENG J, et al. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks [J]. Algorithms, 2016, 9(2):41.
- [55] NI J, MA X, XU L. An image recognition method based on multiple BP neural networks fusion[C]// International Conference on Information Acquisition. 2004:323-326.
- [56] ZHANG J X, LI S Y. Air quality index forecast in Beijing based on CNN-LSTM multi-model[J]. Chemosphere, 2022, 308(1):136-180.
- [57] CHENG D L, ZHANG D W, CHEN Y X. A review of multimodal emotion recognition[J]. Journal of Southwest University for Nationalities(Natural Science Edition), 2022, 48(4):440-447.
- [58] MU Z J, FU Y R. A review of multimodal learning analysis[J]. Modern Educational Technology, 2021, 31(6):23-31.
- [59] ZHOU X M, HU Y G, LIU W J, et al. Research on Urban function Identification based on multi-modal and multi-level data fusion method[J]. Computer Science, 2021, 48(9):50-58.
- [60] LIU J, LI T, XIE P. Urban big data fusion based on deep learning: an overview[J]. Information Fusion, 2020, 53:123-133.
- [61] RAMACHANDRAM D, TAYLOR G W. Deep multimodal learning a survey on recent advances and trends[J]. IEEE Signal Processing Magazine, 2017, 11(13):96-108.
- [62] VIELZEUF V, LECHERVY A, PATEUX S. Centralnet: a multilayer approach for multimodal fusion[C]// Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018.
- [63] LIU S, WANG G R, LI Y L, et al. A review of Chinese short text classification techniques[J]. Journal of Information Engineering University, 2021, 22(3):304-312.
- [64] CHO K, MERRIENBOER B V, GULCEHRE C. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics. 2014:1724-1734.
- [65] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]// Proceedings of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014:3104-3112.
- [66] MURPHY R R. Computer vision and machine learning in science fiction[J]. Science Robotics, 2019, 4(30).
- [67] YANG Z, HE X, GAO J. Stacked attention networks for image question answering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:21-29.
- [68] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2012:1106-1114.
- [69] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]// Proceedings of Annual Conference on Neural Information Processing Systems. Cambridge, MA: MIT Press, 2014:568-576.
- [70] LEE W, KIM J, LEE N K. Pornographic Video Detection with Convolutional Two-Stream Network Fusion[C]// 11th International Conference on Information and Communication Technology Convergence (ICTC). 2020:1273-1275.
- [71] JIANG X, WU F, ZHANG Y. The classification of multi-modal data with hidden conditional random field[J]. Pattern Recognition Letters, 2015, 51:63-69.
- [72] NEFIAN A V, LIANG L, PI X. Dynamic Bayesian networks for audio-visual speech recognition [J]. EURASIP Journal on Advances in Signal Processing, 2002(11):1-15.
- [73] SHI Q P. Audiovisual speech recognition system based on HMM [D]. Nanjing: Hehai University, 2011.
- [74] LIU Y, LIU L, GUO Y. Learning visual and textual representations for multimodal matching and classification [J]. Pattern Recognition, 2018, 84(12):51-67.
- [75] BALTRUŠAITIS T, AHUJA C, MORENCY L P. Multimodal machine learning: A survey and taxonomy[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2018, 41(2):423-443.
- [76] ZADEH A, LIANG P P, MAZUMDER N. Memory fusion network for multi-view sequential learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018.
- [77] HUANG P S, HE X, GAO J. Learning deep structured semantic models for web search using click-through data [C]// Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. 2013:2333-2338.
- [78] WU A, HAN Y. Multi-modal circulant fusion for video-to-language and backward[C]// International Joint Conference on Artificial Intelligence. 2018.
- [79] WANG J X, FANG S N, FENG Y X, et al. Design and implementation of intelligent police protection platform based on big data[J]. Police Technology, 2021(3):39-42.
- [80] XU H M, DENG H C. Research on key technologies and applications of government big data Center construction[J]. Smart Buildings and Smart Cities, 2022, (7):94-96.
- [81] LI G L, ZHOU X H. An overview of AI-oriented data management techniques[J]. Journal of Software, 2021, 32(1):21-40.



YAN Jiahe, born in 1994, Ph.D candidate. Her main research interests include data mining and data fusion.



LI Honghui, born in 1964, professor, Ph.D supervisor. Her main research interests include software testing, big data governance and rail transit information technology.