智慧城市多源异构大数据处理框架

刘岩[1]，王华[2]，秦叶阳[3]，朱兴杰[1]
1. 泰康保险集团股份有限公司数据信息中心，北京 102206；
2. 中国人民大学，北京 100872；
3. 北京大学，北京 100871

摘要
智慧城市建设的重心已由传统IT系统和信息资源共享建设，转变为数据的深度挖掘利用和数据资产的运营流通。大数据中心是数据资产管理和利用的实体基础，其核心驱动引擎是大数据平台及各类数据挖掘与分析系统。讨论了智慧城市大数据中心建设的功能架构，围绕城市多源异构数据处理的实际需要，对数据中心大数据平台的架构进行了拆分讲解，并以视频大数据处理为例，阐述了数据中心中大数据平台的运转流程。

关键词
智慧城市；大数据；多源异构；视频分析

中图分类号：TP391　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2017007

Multisource heterogeneous big data processing architecture in smart city

LIU Yan[1], WANG Hua[2], QIN Yeyang[3], ZHU Xingjie[1]
1. Data & Information Services Center, Taikang Insurance Group Co., Ltd., Beijing 102206, China
2. Renmin University of China, Beijing 100872, China
3. Peking University, Beijing 100871, China

Abstract
The focus of smart city construction has been transferred from the tradition IT systems and sharing of information resources construction into the data mining operations and the flow of data assets. Big data center is the physical infrastructure of data assets management and utilization. Its core driver includes big data platform and kinds of data mining and analysis systems. The functional architecture of big data center in smart cities was discussed. And around the actual needs of urban multisource heterogeneous data processing, the structure of the big data platform used by parts was explained. Then taking the video processing as an example, the working flow of big data platform in the big data center was described.

Key words
smart city, big data, multisource and heterogeneous, video analysis

1 引言

随着智慧城市建设逐步由信息基础设施和应用系统建设迈入数据资产集约利用与运营管理阶段，城市大数据中心已成为智慧城市打造核心竞争力、提升政府管理效能的重要工具。一方面政府借助大数据中心建设可以将有限的信息基础设施资源集中高效管理和利用，大幅降低各自为政、运维机关庞杂、财政压力过大的问题；另一方面，可以在国务院、发展和改革委员会大力支持的政策东风下，打破部门间数据壁垒，推动政府各部门职能由管理转为服务，提高数据共享利用率和透明度。以大数据中心为核心构建城市驾驶舱，实现城市运转过程的实时全面监控，提高政府决策的科学性和及时性。智慧城市大数据中心建设功能框架如**图1**所示，其中针对不同部门的数据源，由数据收集系统完成数据的汇聚，并根据数据业务类型和内容的差异进行粗分类。为避免过多“脏数据”对大数据平台的污染，对于批量数据，不推荐直接将数据汇入大数据平台，而是单设一个前端原始数据资源池，在这里暂时存储前端流入的多源异构数据，供大数据平台处理调用。

大数据平台是城市大数据中心运转的核心驱动引擎，主要完成多源数据导入、冗余存储、冷热迁移、批量计算、实时计算、图计算、安全管理、资源管理、运维监控等功能[1]，大数据平台的主体数据是通过专线连接或硬件复制各政府部门数据库的方式获得，例如地理信息系统（geographic information system，GIS）数据、登记信息等。部分数据通过直连业务部门传感监测设备的方式获得，例如监控视频、河道流量等。大数据平台的输出主要是结构化关联数据以及统计分析结果数据，以方便各类业务系统的直接使用。

不同部门间共享与交换的数据不推荐直接使用原始数据，一方面是因为原始数据内容密级存在差异，另一方面是因为原始数据内容可能存在错误或纰漏。推荐使用经过大数据平台分类、过滤和统计分析后的数据。不同使用部门经过政务信息门户统一需求申请和查看所需数据，所有数据的交换和审批以及数据的监控运维统一由数据信息中心负责，避免了跨部门协调以及数据管理不规范等人为时间的损耗，极大地提高了数据的流通和使用效率。另外，针对特定的业务需求，可以基于大数据平台拥有的数据进行定制开发，各业务系统属于应用层，建设时不宜与大数据平台部署在同一服务器集群内，并且要保证数据由大数据平台至业务系统的单向性，尽量设置业务数据过渡区，避免应用系统直接对大数据平台核心区数据的访问。

目前主流大数据平台都采用以Hadoop为核心的数据处理框架，例如Cloudera公司的CDH（Cloudera Distribution for Hadoop）和星环信息科技（上海）有限公司（Transwarp）的TDH（Transwarp Data Hub）、Apache Hadoop等。以Hadoop为核心的大数据解决方案占大数据市场95%以上的份额，目前国内80%的市场被 Cloudera占有，剩余20%的市场由星环信息科技（上海）有限公司、北京红象云腾系统技术有限公司、华为技术有限公司等大数据公司分享。随着数据安全意识的增强、价格竞争优势的扩大，国内企业在国内大数据市场的份额和影响力正在快速提升。大数据的应用历程可归纳为3个阶段：第一个阶段是面向互联网数据收集、处理的搜索推荐时代；第二个阶段是面向金融、安全、广播电视数据的用户画像和关系发现时代；

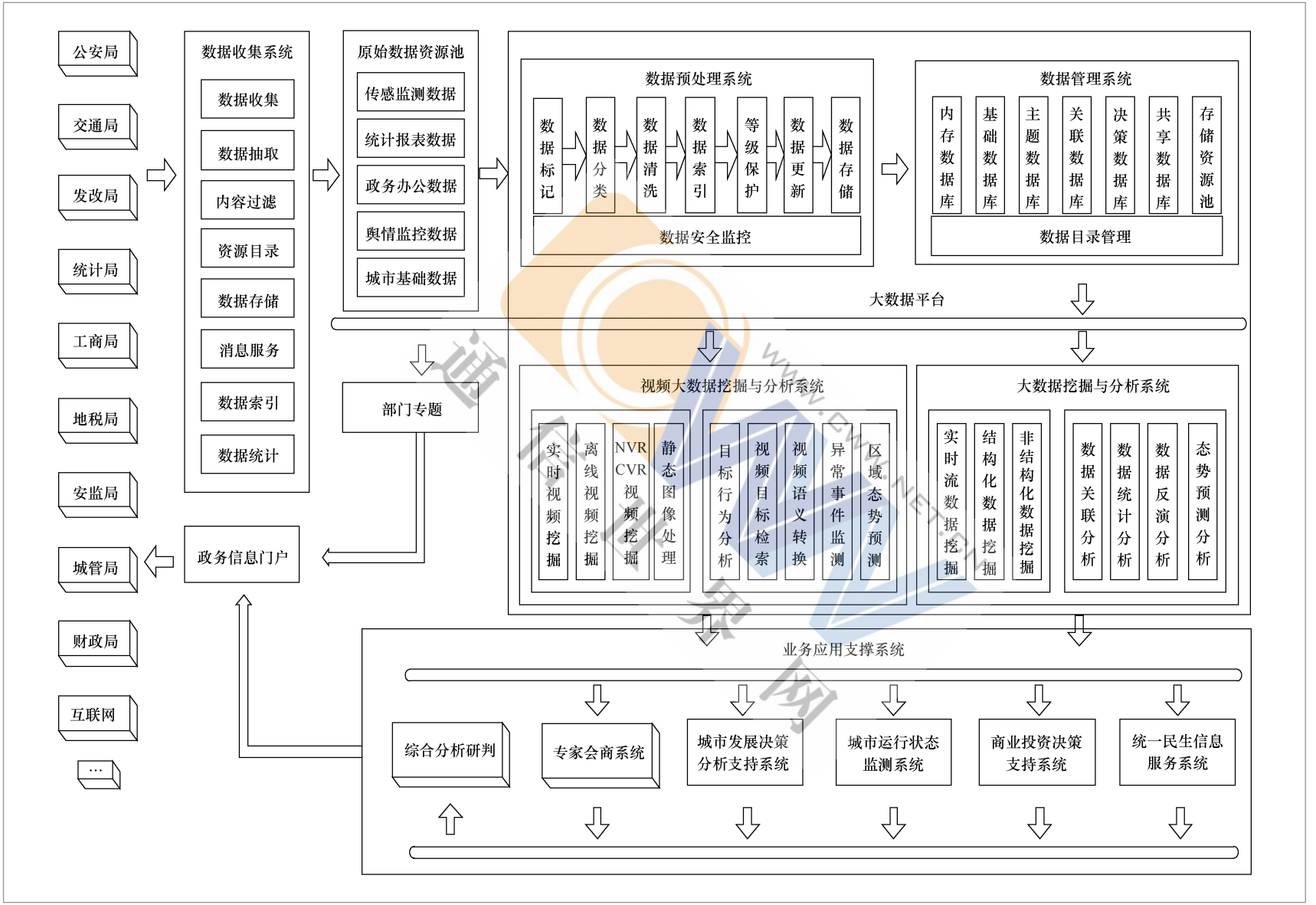


图1 智慧城市大数据中心功能框架

第三个阶段是面向多数据源与多业务领域数据的融合分析与数据运营时代，并且对数据处理规模和实时性的要求大幅提高。

本文在智慧城市大数据中心建设方案的基础上，阐述了多源异构大数据处理的框架和流程，并以最典型的非结构化视频大数据处理为例，介绍了多源异构大数据处理框架运转的流程。

2 多源异构大数据处理框架

2.1 系统整体架构

多源异构是大数据的基本特征[12]，为适应此类数据导入、存储、处理和交互分析的需求，本文设计了如图2所示的系统框架，主要包括3个层面的内容：基础平台层、数据处理层、应用展示层。其中，基础平台层由Hadoop生态系统组件以及其他数据处理工具构成，除了提供基本的存储、计算和网络资源外，还提供分布式流计算、离线批处理以及图计算等计算引擎；数据处理层由多个数据处理单元组成，除了提供基础的数据抽取与统计分析算法外，还提供半结构化和非结构化数据转结构化数据处理算法、数据内容深度理解算法等，涉及自然语言处理、视频图像内容理解、文本挖掘与分析等，是与人工智能联系最紧密的层，该层数据处理效果的好坏直接决定了业务应用层数据统计分析的准确性和客户体验；应用展示层由SSH（Struts+Spring+Hibernate）框架及多类前端可视化工具组成，对应用层的约束是比较宽松的，主要是对数据处理层结果的进一步归纳和总结，以满足具体业务的需要。系统框架的使用优先推荐开源生态系统及其组件，系统存储主要依托Hadoop分布式文件系统（Hadoop distributed file system，HDFS）、HBase，同时支持Oracle、MySQL等结构化数据存储系统，计算框架涵盖MapReduce、Storm、Spark以及定制分布式视频流处理引擎，可视化系统基于SSH框架设计，可根据实际需求，灵活配置。

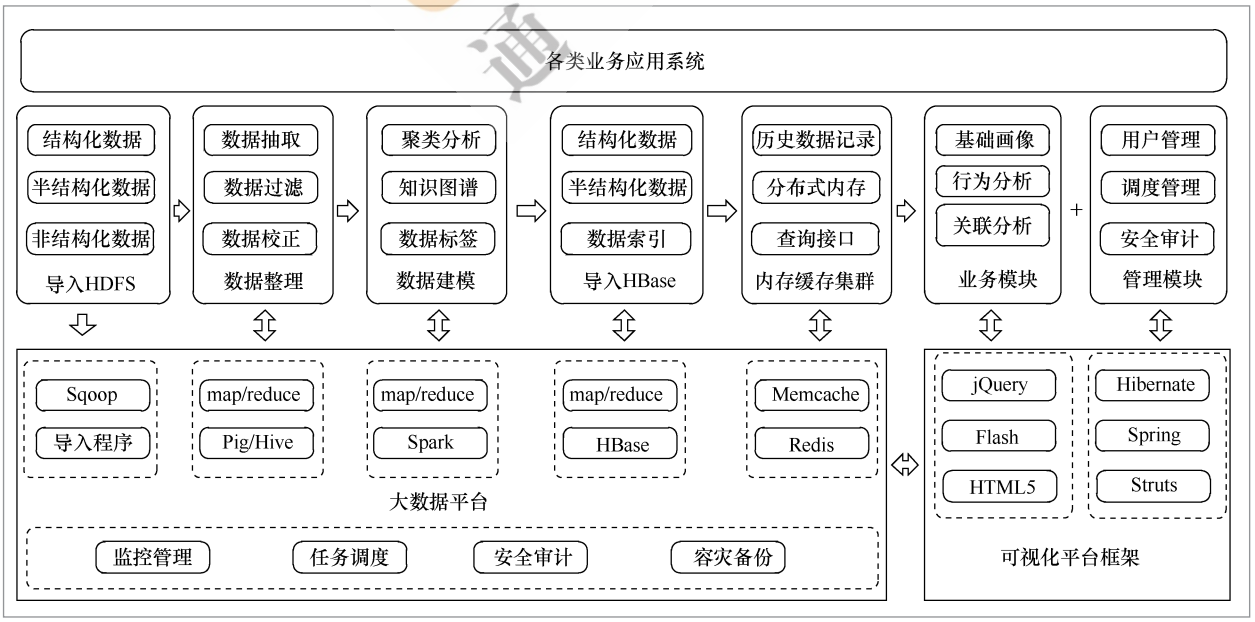


图 2 多源异构大数据处理框架

2.2 多源数据导入

鉴于数据存储介质、数据存储类型和数据传输方式的差异，系统在数据导入单元设计了如下数据导入框架，借助不同的导入工具，实现不同源数据和不同结构数据的导入，如图3所示。其中，对实时性要求较高的监测数据以分布式消息队列的形式由Kafka分发；关系型数据库使用Sqoop等工具，直接将数据导入HDFS[3,4]；对于安全等级较高的数据和其他一些离线数据，使用硬件复制或文件传输协议（file transfer protocol，FTP）传输的方式导入；对于日志等文本数据使用Flume工具导入；对于互联网数据使用爬虫程序爬取，并导入；对于视频等多媒体数据，使用各厂商提供的定制码流软件开发工具包（software development kit，SDK）开发导入程序，或者利用多媒体流处理引擎直接抓取和在线处理。在智慧城市建设过程中，数据来源差异一般较大，数据库中存放的主要是经过业务系统加工后的数据，而描述行为过程的数据一般都未被记录，此时，需要定制开发能够直接连接原始数据源的数据采集工具。

2.3 异构数据处理

根据数据类型的差异，选择不同的计算和存储引擎。对于非实时性数据计算，选择MapReduce计算引擎[5]；对实时性要求较高的数据计算，选择Spark或Storm计算框架[6,7]；对时序不可分的流媒体数据处理，选择定制流媒体计算引擎，如图4所示。对于结构化或键值对数据，采用Hive或HBase存储，兼容Oracle和MySQL等关系型数据库；对于日志、多媒体等半结构化和非结构化数据，采用HDFS存储。数据仓库可以统一建立在HDFS上，统一的存储有助于最大化地发挥分布式系统的数据处理能力，充分利用内网带宽，减少异构数据仓库自身性能瓶颈导致的大数据系统性能下降问题。

对于结构化数据的处理主要包括内容清洗、统计分析、关联分析等；对于半结构化数据的处理涉及模板分类、字段检索、关键字段提取等；对于非结构化数据的处

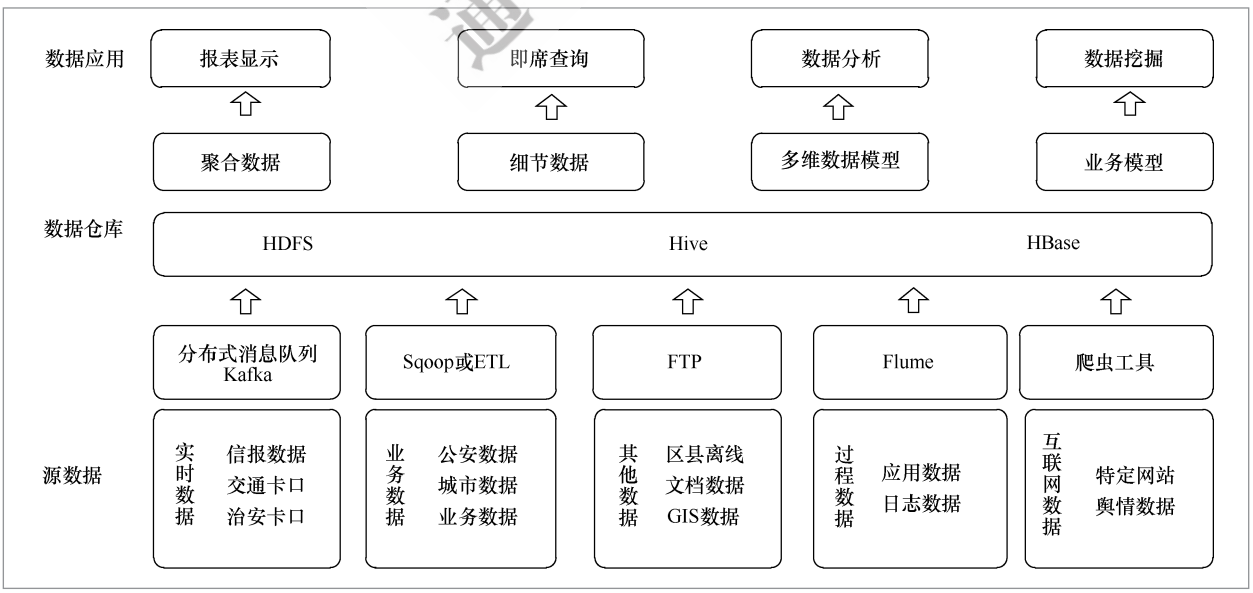


图 3 多源数据导入框架

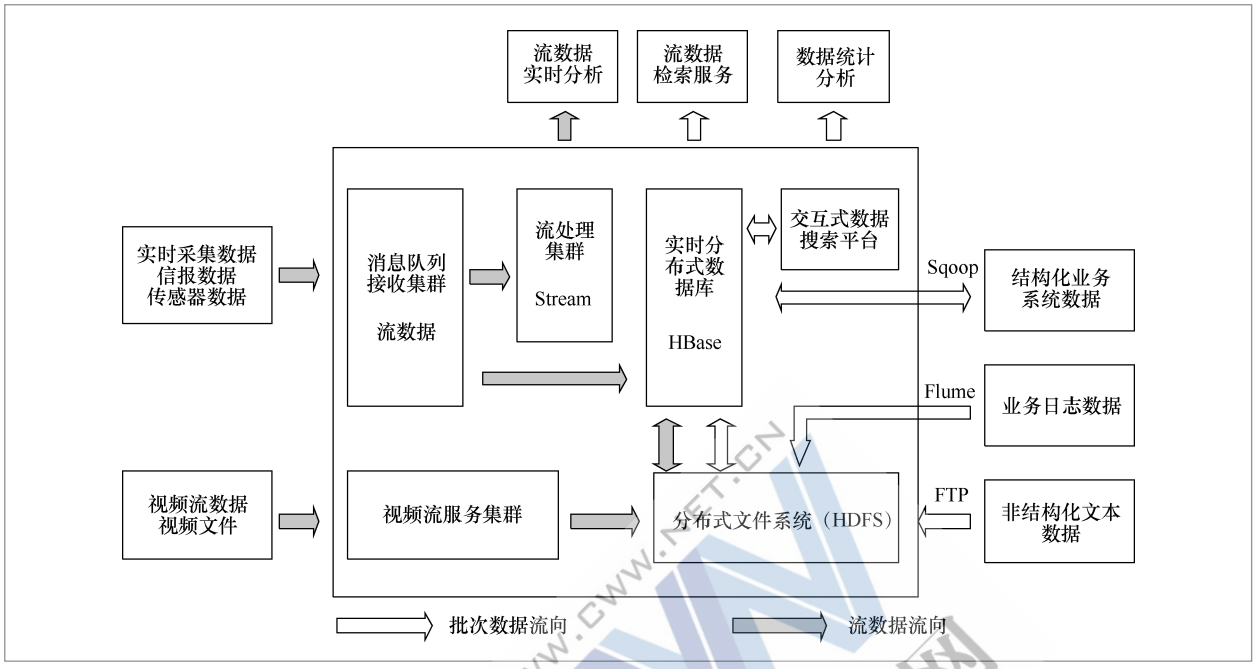


图 4 异构数据处理框架

理涉及音视频内容的结构转化、文本内容的挖掘与分析、语义理解与情感分析等。随着数据结构多样性和内容不确定性的增加，数据处理的复杂度和难度呈现指数型非线性增长，诸多数据处理问题在这个阶段转变为人工智能算法问题。

2.4 统一运维管理

大数据平台的运维管理借助统一运维管理平台实现，管控平台具备大数据平台定制化组件安装、资源灵活配置、字段级权限控制、账户管理等功能，借助统一的运维管理平台，对平台安装节点的CPU、内存、硬盘资源进行控制，并对节点所在机架进行规划，通过运维管理主节点，可实现大数据平台的自动部署和安装，与此同时，运维管理平台可实时监控正在运行的各服务的资源使用情况和任务进度情况，为各服务提供资源隔离或资源抢占式两种选择方案，灵活配置服务运行节点，大大节省运维管理人员的工作量。

3 视频数据处理应用示例

在智慧城市建设中，视频不仅是存储规模最大的数据，同时也是最典型的异构大数据，数据内容在不同的处理阶段，表现为不同的数据形式：非结构化（视频、图像）、半结构化（特征点）、结构化（特征向量、描述属性）。视频数据[8]不仅用于治安侦查、违章监测，还被用于城市人群密度监测，结合舆情、地理定位等信息，可用于对城市不同区域安全等级的评估。视频数据处理算法框架如图5所示，视频数据处理的过程是逐步将非结构化数据转为结构化数据，然后做统计和关联分析的过程。

3.1 视频数据标记

视频数据标记有助于提高视频内容提取和描述的准确性和稳定性，使得视频内容检测与分析算法的设计更有针对性，原

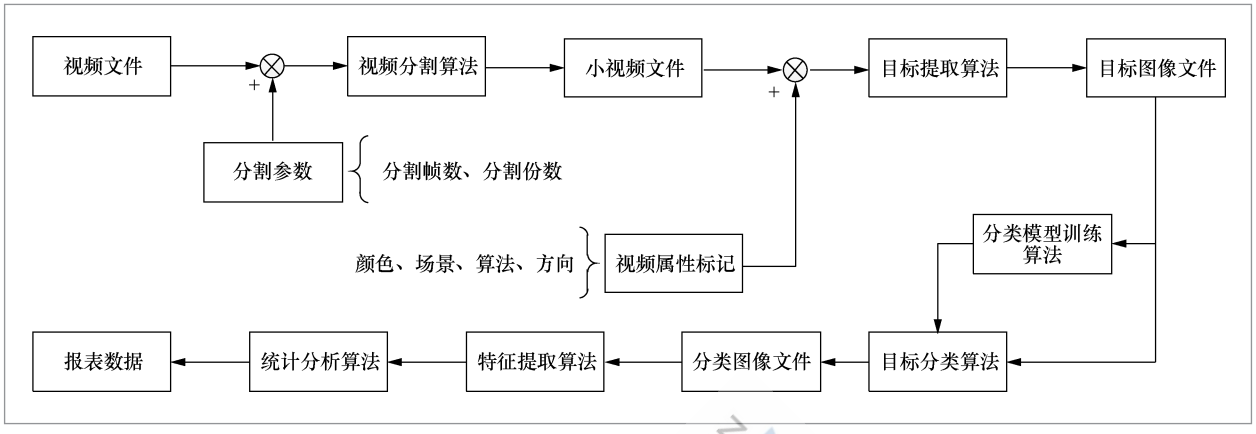


图 5 视频数据处理流程

则上对视频内容的结构化描述信息越全面越好，但是容易受标记人员主观因素干扰，因此只选择容易区分和定义的以下几类标记信息：视频场景、视频主色、运动方向、适用算法。视频场景可分为：卡口、路口、广场、街道等，视频主色可分为：彩色和灰色，运动方向根据图像坐标系分为8个方向，适用算法主要用于标记该视频适用于哪类算法，例如行人检测、遗留物检测、交通标志检测、车牌检测等。标记后的视频经过视频分割算法处理，被切分成大小适合MapReduce处理的文件块。

3.2 视频内容挖掘

视频多媒体数据包含的信息非常丰富，这里仅以视频中的人、车、自行车目标的检测与跟踪为例，阐述非结构化视频大数据内容挖掘的实现过程。

视频内容挖掘是通过对视频文件或视频流的解码，逐帧进行分析处理的。视频中的运动目标是检测的主要对象，通过背景建模、前景目标分割算法确定潜在运动目标的位置，然后通过运动目标跟踪算法对粘连目标、误分割目标以及特征不稳定目标进行切分、合并和过滤处理，处理流程如图6所示，图6中对不同的运动目

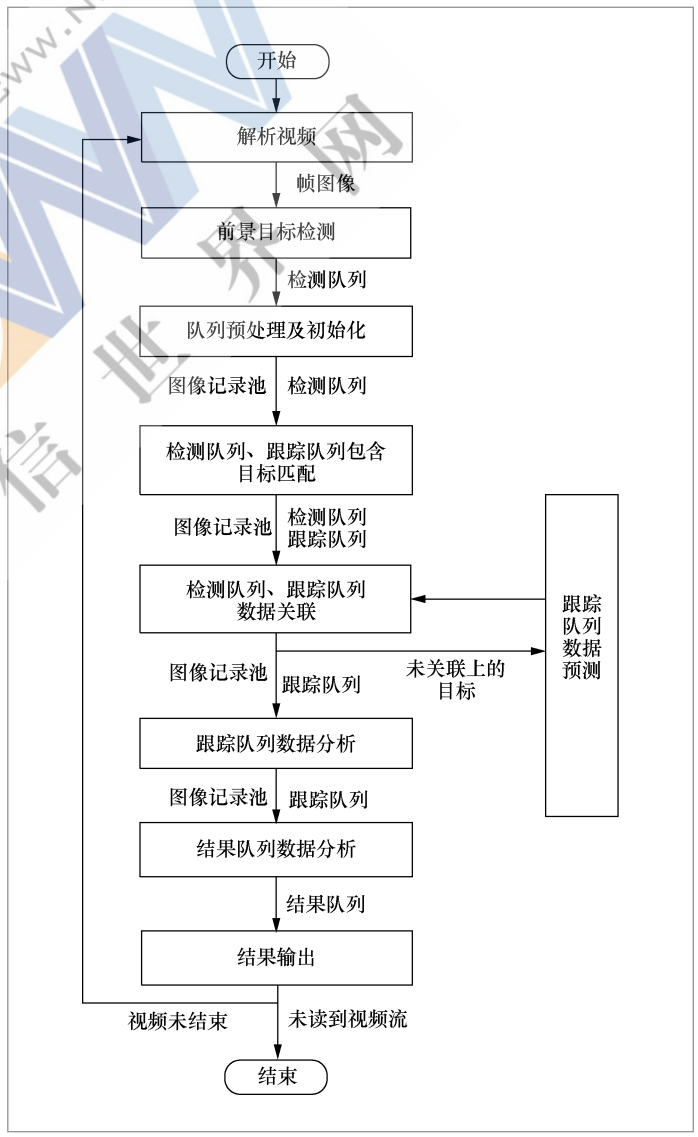


图 6 视频内容挖掘流程

标分别建立检测存储队列、跟踪存储队列、结果存储队列，用以实现基于视频前后帧序列的目标过滤与判定。整个处理过程可以使用MapReduce框架实现，值得注意的是，视频对象处理需要耗费大量的内存资源，单靠Java虚拟机（Java virtual machine，JVM）已难以满足需求，因此，推荐使用C+Java的混合语言编程处理模式。

3.3 视频目标分类

对视频内容挖掘单元输出的目标图像文件做进一步显著性检测与分类判定，主要包括图像中的人体检测、车辆检测、自行车检测，并对目标图像中包含多个目标的情况进行切分，对误检或位置不精确的目标进行过滤或校正。

本文使用优化的弹性形变模型（deformable parts model，DPM）算法对目标图像进行二次检测，如图7所示。为提高检测精度，对尺寸（宽或高）小于320像素的图像进行插值处理，扩大至（宽或高）640像素，二次检测的结果仍以图像文件的形式存储在HDFS上，文件属性及其与原视频流的对应关系记录在HBase中，该对应关系主要包括原视频路径、图像对应视频中的帧序号等。

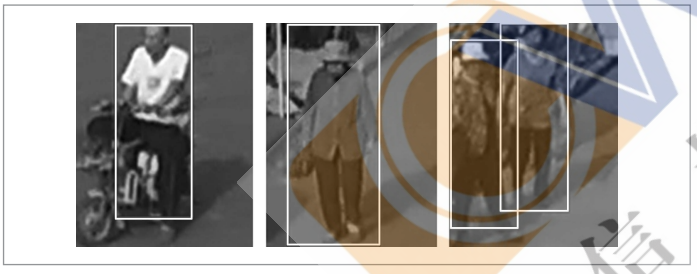


图 7 基于优化 DPM 的行人二次定位示例

3.4 视频目标检索

视频目标检索是在视频目标分类结果的基础上，对图像内容进行结构化特征描述[9]，特征向量冷数据存储在HBase中，热数据存储在内存中，每一次的检索查询是对所有图像数据特征的相似性比较。其中特征向量的构建综合考虑颜色不变性和尺度不变性的现实需求，使得特征向量对颜色变化敏感而对尺度变化顽健，目标间的相似性通过特征向量余弦计算。视频监控目标检索示例如图8所示。

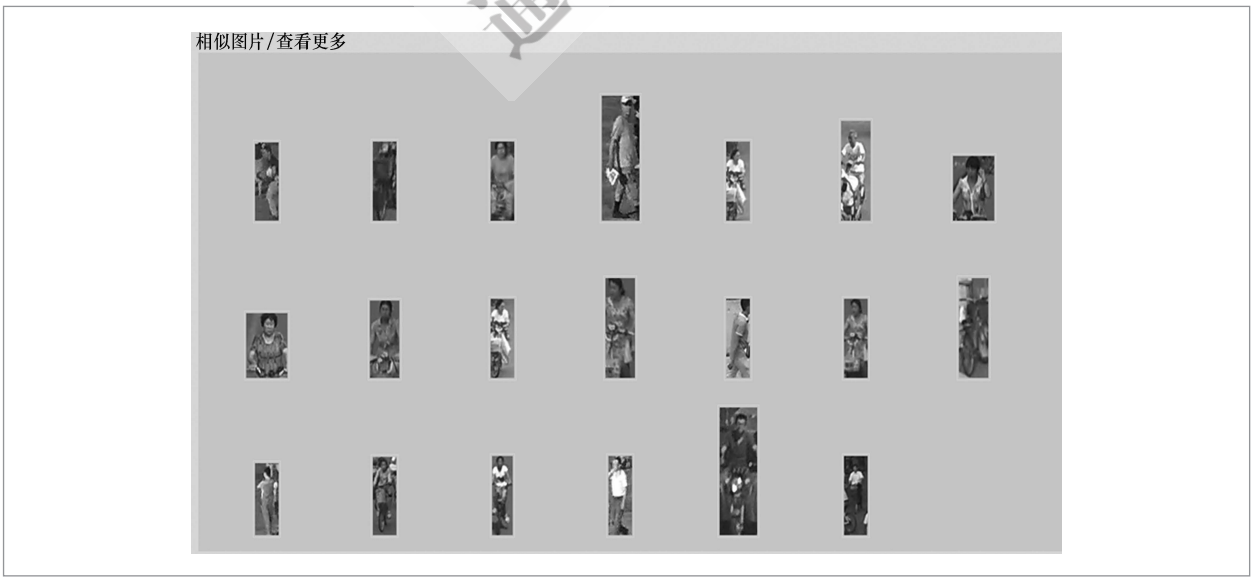


图 8 视频监控目标检索示例

3.5 区域密度监测

如图9所示，将检测到的人、车、自行车等以行为人为主体的目标与监控摄像机的地理位置结合在一起，得出人车分布情况和城市活跃度情况。图9（a）以曲线形式展示了不同时刻的人车分布情况，图9（b）为基于密度波动的城市活跃度评分。

4 结束语

在智慧城市建设中，大数据中心扮演着城市大脑的角色，汇聚了来自不同业务部门、不同企事业单位和不同行为人的过程、行为和位置等数据，这些城市主体元素的监测数据组成了大数据中心庞杂的数据源，大数据平台及各类数据挖掘与分析系统组成了大数据中心的数据分析引擎。在政府角色由城市管理转向城市运营和服务的过程中，大数据中心建设起到了重要的推动作用。本文从智慧城市大数据中心运转的角度，介绍了大数据中心对多源异构大数据处理的架构体系，并且以最典型的视频大数据处理为例，讲解了大数据平台中非结构化数据处理的方法和流程，最后给出了数据挖掘结果如何服务于智慧城市的应用示例。

参考文献:

[1] 程学旗，靳小龙，王元卓，等. 大数据系统和分析技术综述[J]. 软件学报，2014，25(9)：1889-1908.
CHENG X Q，JIN X L，WANG Y Z，et al. Survey on big data system and analytic technology[J]. Journal of Software，2014，25(9)：1889-1908.

[2] 石宇，詹明，尹璐，等. 面向对象的多源异构数据关联组织与分析[J]. 测绘通报，2015(1)：102-104.
SHI Y，ZHAN M，YIN L，et al. Research on associated orgabization and analysis of target-oriented multi-source heterogeneous data[J]. Bulletin of Surveying and Mapping，2015(1)：102-104.

[3] GHEMAWAT S，GOBIOFF H，LEUNG S. File and storage systems：the Google file system[J]. ACM Sigops Operating Systems Review，2003，37(5)：29-43.

[4] HE H，DU Z，ZHANG W，et al. Optimization strategy of Hadoop small file storage for big data in healthcare[J]. Journal of Supercomputing，2015，72(10)：1-12.

[5] DEAN J，GHEMAWAT S. MapReduce：simplified data processing on large clusters[J]. Communications of the ACM，2008，51(1)：107-113.

[6] 孙大为，张广艳，郑纬民. 大数据流式计算：关键技术及系统分析[J]. 软件学报，2014，25(4)：839-862.

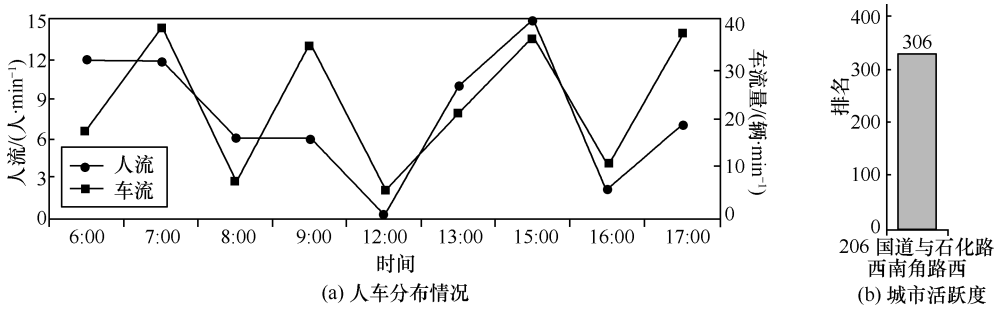


图 9 城市区域密度监测示例

SUN D L, ZHANG G Y, ZHENG W M. Big data stream computing: technologies and instances[J]. Journal of Software, 2014, 25(4): 839-862.

[7] 齐开元, 赵卓峰. 针对高速数据流的大规模数据实时处理方法[J]. 计算机学报, 2012, 35(3): 477-490.
QI K Y, ZHAO Z F. Real-time processing for high speed data stream over large scale data[J]. Chinese Journal of Computers, 2012, 35(3): 477-490.

[8] DING S H, LI G, LI Y, et al. SurvSurf: human retrieval on large surveillance video data[J]. Multimedia Tools & Applications, 2016(1): 1-29.

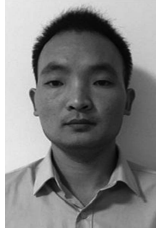
[9] ZHU H D, SHEN Z, SHANG L, et al. Parallel image texture feature extraction under Hadoop cloud platform[J]. Springer International Publishing, 2014(8588): 459-465.

作者简介

刘岩(1982-),男,泰康保险集团股份有限公司数据信息中心高级工程师、高级主管,中国计算机学会会员,主要研究方向为智慧城市建设与规划、多源异构大数据内容挖掘与分析、人工智能理论与应用等,在大数据系统设计、人脸识别、OCR识别等领域具有丰富的实践经验,曾作为首席专家参与多个城市智慧化发展规划与实施建设。目前已发表学术论文25篇,申请美国发明专利4项,中国发明专利17项,软件著作权3项,荣获省科技进步奖一项,承担多个“973”计划项目、国家自然科学基金等项目。

王华(1985-),男,中国人民大学硕士生,主要研究方向为大数据处理架构与应用、多源异构数据内容清洗及结构化转化等,对Hadoop、Spark生态系统及组件具有丰富的应用实践经验。

秦叶阳(1986-),女,就职于北京大学,安徽荣创智能科技有限公司联合创始人,主要研究方向为智慧城市信息化建设、大数据处理系统设计与应用、信息安全等,在信息化系统建设、项目组织与运营管理、公共关系管理等方面具有丰富的经验。



朱兴杰(1986-),男,泰康保险集团股份有限公司数据信息中心应用创新高级工程师,主要研究方向为视频数据内容挖掘与分析、人脸检测与识别、机器学习等。

收稿日期: 2016-11-18