

**大数据的集成与融合**

**组员：潘恒禧**

**郭新镜**

**叶达豪**

**钟文丽**

**大数据的集成与融合**

**摘要：**大数据，人工智能时代下,每天都产生各种各样的数据。这些数据中有结构化的，半结构化的，非结构化的数据，比如微博留言，图片，视频上传，物联网，电商购物，等等。然而，将这些数据集成与融合成为了数据价值挖掘的关键一步。通过数据的集成与融合我们能够更深度地挖掘数据的价值。本文主要是综述地讲述大数据的融合与集成。首先，从数据融合的起源，到数据仓库，再到基于Hadoop的大数据融合架构讲述数据融合与集成的发展。其次，新时期下，一方面人工智能迅速发展，人工智能下的传感器的数据融合是当下数据融合的关键，另一方面，数据仓库，基于Hadoop的大数据融合架构更多处理的是结构化的数据，大数据时代下非结构化的数据(文本，图片，视频等等)是海量的，文章探索了多模态数据(文本，图片，视频等等)的融合。最后，通过寻找实际应用案例，利用这些案例来对我们所研究的技术进行一个很好的描述。此外，通过案例的分析，我们可以延伸到未来大数据集成和融合技术的发展趋势，让我们拥有更为前瞻的视野去面对未来的技术发展。

**关键词：**数据融合；人工智能；大数据

Integration and integration of big data.

**Abstract**--- Big data, under the age of artificial intelligence, produce all kinds of data every day.These data have structured, semi-structured, unstructured data, such as tweets, pictures, video uploads, Internet of things, e-commerce shopping, and so on.However, integration and integration of these data is a key step in data value mining.Through the integration and integration of data, we can excavate the value of data more deeply.This article mainly describes the integration and integration of big data.First of all, from the origin of data fusion, to the data warehouse, to the big data fusion architecture based on Hadoop, the development of data fusion and integration.Second, under the new period, on the one hand, rapid development of artificial intelligence, artificial intelligence of sensor data fusion is the key data fusion, on the other hand, data warehouse, large data fusion based on Hadoop architecture more processing is structured data, the era of big data unstructured data (text, pictures, video, etc.) is the mass, the article explores the multimodal data (text, pictures, video, etc.).Finally, by looking for practical application cases, these cases are used to give a good description of the technology we are studying.In addition, through the analysis of cases, we can extend the development trend of big data integration and fusion technology in the future, so that we can have a more forward-looking vision to face the future technological development.

**Keywords:** Data fusion;Artificial intelligence;Big data

**目录**

[1. 绪论 1](#_Toc505071692)

[1.1研究背景及意义 1](#_Toc505071693)

[1.2研究现状 2](#_Toc505071694)

[1.3论文组织结构 2](#_Toc505071695)

[2.数据融合相关理论与技术 3](#_Toc505071696)

[2.1数据融合的基本概念及发展 3](#_Toc505071697)

[2.2数据融合的级别 4](#_Toc505071698)

[2.3数据仓库技术ETL 5](#_Toc505071699)

[3. 大数据时代数据融合与集成 7](#_Toc505071700)

[3.1大数据的特点 7](#_Toc505071701)

[3.2大数据的来源 8](#_Toc505071702)

[3.3大数据的数据抽取和集成需求 9](#_Toc505071703)

[3.4大数据的数据融合与集成架构 10](#_Toc505071704)

[3.5比较传统面向数据仓库和大数据集成的各自目的和特点 13](#_Toc505071705)

[4. 大数据时代数据融合与集成趋势 14](#_Toc505071706)

[4.1人工智能下传感器数据集成与融合 14](#_Toc505071707)

[4.1.1多传感器数据融合面对的问题 16](#_Toc505071708)

[4.1.2传感器资源信息元数据模型 16](#_Toc505071709)

[4.1.3多传感器融合方法 17](#_Toc505071710)

[4.2大数据下非结构化(多模态)数据的融合 19](#_Toc505071711)

[4.2.1特征提取 21](#_Toc505071712)

[4.2.2深度网络隐藏层特征提取 24](#_Toc505071713)

[4.2.3隐藏层特征融合 26](#_Toc505071714)

[5. 案例分析 28](#_Toc505071715)

[5.1智慧城市的数据融合与集成应用 28](#_Toc505071716)

[5.2基于多模态数据融合的车辆跟踪应用 30](#_Toc505071717)

[5.3基于多模态数据融合的图文检索应用 33](#_Toc505071718)

[6. 总结与数据融合技术的展望 35](#_Toc505071719)

[6.1总结 35](#_Toc505071720)

[6.2数据融合技术的展望 35](#_Toc505071721)

[参考文献： 36](#_Toc505071722)

# 

# 1. 绪论

## 1.1背景及意义

大数据（big data）概念最早是在 20 世纪 80 年代提出的，2008年9月，《科学》杂志发表文章“Big Data：Science in the Peta byte Era”，“大数据”开始广泛传播。随着在大数据时代的到来，几乎每个行业都赶上了大数据、人工智能、机器学习的潮流，高速发展的互联网给我们带来很多突破和进步，与此同时也面临着来自不同领域的、复杂多样的数据集，处理信息的难度就大大增加了。数据融合一词最早出现在20世纪70年代末，用于军事领域，开展声呐信号解释系统的研究。数据融合技术可以避免单一传感器的局限性，避免数据来源的单一性，提高结论的可信度，获得并推理出更多的信息。

作为前沿领域技术，无论是军用系统还是民用系统，目前都趋向于采用数据融合技术来进行信息综合处理。在知识爆炸的信息时代，数据融合的工作就变得尤为重要，做得不好将会导致数据富有，但信息贫乏的情况发生。它将在多传感器数据冲突处理，智能控制，决策支持系统、工业设备监测和生产自动化、战场多目标精确跟踪与识别、GPS导航、机器人控制、自动故障诊断以及Internet管理等诸多领域有着广阔的发展前景和巨大的应用价值。

应用数据融合技术具有如下优点：

（1）可以提高信息的可信度。能更加准确地获得环境目标的某一特征或一组相关特征，使整个系统所获得的综合信息具有更高的精度及可靠性。

（2）减少了获得信息的代价。在相同的时间内获得更多的信息，特别是在测量运动速度快的目标时，从而减少了获得同样多信息的代价。

（3）减少信息获取事件，加快信息处理速度，提高信息再用率。由于传感器信息处理是并行进行的，各个单独的传感器可以相对简化其处理步骤，加之计算机技术在数据融合中的大量应用，许多原来需压缩的原始数据，可以直接作为数据融合系统的输入，通过多组这种数据的互相关联，最大限度地利用其中的信息，减少系统信息处理的总时间。

（4）提高了系统容错能力。由于多个传感器所采集的信息具有冗余性，当系统中有一个甚至几个传感器出现故障时，尽管某些信息容量减少了，但仍可由其他传感器获得有关信息，使系统继续运行，故经过信息融合处理无疑会使系统在利用这些信息时具有很好的容错性能。

## 1.2现状

数据融合是一个新兴的研究领域，近年来，数据融合技术不断应用到各个行业中，取得了一定的效果，包括复杂工业过程控制，机器人。自动目标识别、交通管制、惯性导航、海洋监视和管理、农业、遥感、医疗诊断、图像处理、模式识别等领域。在加拿大，从1991年起，Lockeed Martin研究小组开发和展示了为巡逻护卫舰和飞机做决策的数据融合系统，已经历了四个阶段，使决策过程不断得到优化，利用数据融合技术研制分布式数据融合系统，使目标跟踪性能大大增强，确定目标位置更准确，速度更快。在我国，地质矿产研究人员利用多源数据融合技术对遥感和地学数据信息进行综合分析提高寻矿效果，应用影像融合技术实现地质目标细节与色彩纹理结构的结合，通过信息提取和综合分析进行矿产资源评价，研究地理数据，测量数据和矿物数据的集成，建立可视化应用系统。在交通运输领域则把数据融合技术应用于车辆定位、车辆身份识别、车辆跟踪与导航以及交通管制，实现控制中心对多路数据的采集和综合分析。在城市管理领域，涉及到多方面的数据，利用数据融合技术对航空遥感和航天遥感数据、光谱图像和遥感专题图像、遥感数据和非遥感数据，遥感图像与地图等进行融合，使城市管理更加科学。

可见，数据融合技术，在解决探测、跟踪和目标识别、交通管理等问题方面更加有效。因此，随着数据融合技术的发展，其应用领域也将不断扩大。

## 1.3论文组织结构

本论文主要是介绍大数据的融合与集成，论文的具体的结构如下：

第一章为绪论。首先介绍了大数据集成与融合的背景及意义，接下来介绍了国内外数据集成与融合的现状，最后介绍论文的组织结构。

第二章是数据融合相关理论与技术。首先介绍数据融合的基本概念，然后介绍数据融合的的级别，最后介绍了数据仓库ETL

第三章为大数据时代数据融合与集成。首先介绍了大数据的特点，大数据的来源，主要是以工业数据来举例，接下来介绍了大数据的融合与集成需求，大数据的融合与集成架构，此架构主要是以Hadoop作为基础的。最后，对数据仓库与hadoop为基础的大数据融合与集成架构进行对比。

第四章是数据融合与集成的趋势。介绍了新时期下，数据融合与集成发展方向。首先介绍了人工智能下，传感器数据的融合与集成，然后介绍了大数据下，非结构化(多模态)数据的融合，这对擅长处理结构化数据的数据仓库ETL技术和hadoop为基础的大数据融合与集成架构是一个更好的补充。

第五章为案例分析。主要是通过智慧城市，车辆跟踪，图文检索这三个案例来叙述数据融合与集成技术是如何应用到实际的生活上。

第六章是总结与展望。本章首先对论文中完成的主要工作进行了总结，并在此基础上阐述了不知之处，进而提出了完善该研究的后续工作。

# 2.数据融合相关理论与技术

## 2.1数据融合的基本概念及发展

在早期，数据融合在军事领域得到广泛的重视和应用，是针对一个系统中使用多个传感器这一问题而展开的一种信息处理的、新的研究方向，所以数据融合也称为传感器融合。这对战场上及时准确地获取各种有用的信息，对战场情况和威胁及其重要程度进行适时的完整评价，实施战术、战略辅助决策与对作战部队的指挥控制，是极其重要的。到目前为止，数据融合也没有确切的定义，普遍认为：数据融合技术是指利用计算机对按时序获得的若干观测信息，在一定准则下加以自动分析、综合，以完成所需的决策和评估任务而进行的信息处理技术。数据融合技术，包括对各种信息源给出的有用信息的采集、传输、综合、过滤、相关及合成，以便辅助人们进行态势/环境判定、规划、探测、验证、诊断。

多源信息是数据融合的加工对象，协调优化是数据融合的核心。进入90年代后，随着传感器技术的迅速发展以及军事领域的强烈需求，数据融合引起了广泛的中重视，人们对数据融合开展了多方面的探讨，在许多领域都有一定程度的应用，并成为全球研究的热点之一，世界各国纷纷开展此项技术的研究与应用。与此同时，有几部数据融合领域的学术专著和译著陆续出版。到了90年代中期，数据融合技术在国内已发展成为多方关注的共性关键技术，出现了许多热门研究方向，许多学者致力于分布检测融合、多传感器综合跟踪于定位、目标识别与决策信息融合等。现在由于各种传感器技术的成熟以及云计算单位的成熟，我们通过融合哥哥领域的大数据，进一步解决城市问题、医疗问题、金融风险等问题。

## 2.2数据融合的级别

数据融合通常分为像素级融合、特征级融合及决策级融合。不同级的融合表明融合之前,数据已经被处理的程度。一个给定的数据融合系统,可能涉及所有三个级别数据的输入。

（1）像素级数据融合:这是最低级的数据融合,可以在像素或分辨单元上进行,这也叫作数据级融合。这些数据包括一维时间序列数据、焦平面数据。对于一种相互匹配的形式,录取的数据没有信息损失,因此像素级多传感器处理提供一种最优决策和识别性能。像美国陆地卫星多光谱数据的像素级处理,可以识别病害庄稼的特征,而单一光谱数据则不可能。但是这种像素级数据融合要求精确的传感器配准和高的传输带宽。

（2）特征级数据融合:在这种情况,单个传感器完成目标探测和特征提取处理。多传感器来的目标报告,或经滤波的轨迹则在目标分类之前组合成多源集成(MSI)轨迹。在特征级数据融合中,各个源提供的特征矢量融合到一个综合的特征矢量中,这种融合是比较简单的。特征级数据融合许多情况下是很实用的。有时在像素级组合多源的等同数据是不可能的或是不必要的。

（3）决策级数据融合:在这种情况,每一个传感器先完成对目标的分类,完整的决策则是同另一个分类判决组合产生。因此,对一个目标来说,在分类之前,至少要有两个传感器同时对它进行探测和分类。

下面对像素、特征、决策级数据融合的优缺点进行了比较。研究表明,对多个相似的传感器宜采用特征级融合,不同类型的(或独立的)传感器,建议采用决策级融合。

1. **像素级**

优点:在信号组合和提取之前,没有单个处理的信息损失

局限：1.要求很好地寄存图像

局限：2.通常要求关于目标的多传感器特征的先验知识

局限：3.传感器紧密耦合

(2) **特征级**

优点：1.分类之前保留特征

2.不必寄存图像

局限：1.有一些信息损失

2.要求高维数特征空间

3.在混合的特征空间要求可解译性

(3) **决策级**

优点：1.高度模块化

2.图像可以不精细寄存

局限：1.可能的信息损失

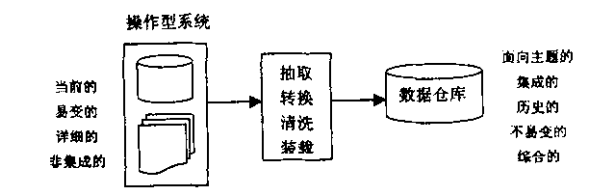
2.如果单个分类器的分类都不可忽略,则决策级融合可能是无效的

## 2.3数据仓库技术ETL

传统数据仓库具有面向主题、集成的、不可更新、随时间而变化等特点，且传统数据挖掘通常处理来自单个域的数据。在大数据时代，我们面对来自不同领域不同来源的数据集的多样性。这些数据集由多种形式组成，每一种形式都有不同的表现、分布、规模和密度。如何从多个完全不同的(但有潜在关联的)数据集中挖掘知识的力量，在大数据研究中是最重要的，从本质上区分大数据和传统数据挖掘任务。这需要先进的技术，可以将知识从不同的数据集有机地融合到机器学习和数据挖掘任务中。数据融合方法可以分为三类:基于阶段的、基于特征层次的数据融合方法和基于语义的基于语义的数据融合方法。最后一种数据融合方法被进一步分为四类:基于多视图的学习型、基于相似度的、基于概率依赖的、基于转移学习型的方法。这些方法侧重于知识融合，而不是模式映射和数据合并，显著区分了跨领域的数据融合和在数据库社区中研究的传统数据融合。

ETL技术是用来描述将数据从来源端经过抽取、转换、加载至目的端的过程，它是构建数据仓库的重要一环，用户从数据源抽取出所需的数据，经过数据清洗,最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。ETL工具的典型代表有:Informatica、Datastage、OWB、微软DTS、Beeload、Kettle等，开源的工具有eclipse的etl插件:cloveretl。

ETL，即数据抽取、转换、清洗、装载的过程，是构建数据仓 库最重要的步骤之一。用户从数据源抽取出所需的数据，经过 数据清洗，最终按照预先定义好的数据仓库模型，将数据加载到数据仓库中去。如图所示。



**图1 ETL**

创建数据仓库最重要的一个步骤是将数据从各种操作型数据库系统中抽取出来，排除数据中的数据缺陷，完成一系列转换，最后将数据加载到数据仓库。这个过程非常繁杂，是整个数

据仓库建设过程中工作量最大的一部分。下面将详细介绍它们的设计细节。

**（l）数据抽取** 并不是源数据库的所有细节数据对于数据仓库的主题域都是有用的，必须根据已确定主题的需要，从原有操作型数据库中抽取相关数据到数据仓库。一般在设计数据

抽取时要考虑以下几个方面：

* 源数据库和目标数据库各自的数据库格式是否一致？
* 从源数据库中要访问哪些文件和表？
* 从源数据库中可以提取哪些字段，抽取记录的条件是什么？
* 目标数据库中的表结构是什么？

**（2）数据转换** 数据仓库中的数据往往来自一个或多个异构的数据库系统，这些数据源之间往往存在着不一致的问题，如不一致的字段长度、不一致的赋值等。数据不一致会严重影响数据仓库的数据质量。数据转换就是处理这些不一致性的过程。数据转换一般包括两个方面的内容。一方面是数据名称及格式的统一，如统一的命名、统一的数据格式、统一的计量单位等等。另一方面，数据仓库中存在着源数据库中可能不存在的数据，因此需要创建新的数据逻辑视图并进行以下的转换：把一个字段的各个部分隔成两个或多个字段。把一个记录的两个或多个字段组合成一个字段。把来自多个记录的字段结合成一个记录。

**（3）数据清洗** 数据质量是决定信息价值的关键因素。高质量的信息导致高质量的决策，而低质量的信息将导致低劣的决策。数据仓库是分析决策的基础，所以数据仓库中数据的准确性是非常重要的。传统数据库中的数据大都存在错误，这是活生生的事实。尽管数据仓库项目能把注意中心放在数据质量问题和引导未来的进一步改进上，但数据仓库开发者却并不能改变数据仓库的基础——历史数据的质量。因此，有必要对进入数据仓库的数据进行全面检查并使他们尽可能无差错。这一过程就称作数据清洗。数据清洗应该处理许多可能存在的错误类型。这些类型包括数据源中丢失数据和有错误数据，还包括两个或多个数据源

里的不一致数据和冲突数据，所有这些都必须处理。有时如果系统不能自动处理，还必须借助于手工操作来完成。

**（4）数据装载** 这一步所进行的工作是运行以上的处理步骤，将数据装入到数据仓库。主要的工作包括确定数据装入的次序、载入初始数据等。目前已有众多商用数据仓库产品走向市场。IBM、OracIe、Sybase、CA、NCR、SAS、Microsoft 等公司已相继推出了各自的数据仓库解决方案，它们的 ETL 工具也各有其优势和不足。在选择 ETL 工具时必须遵守以下原则：支持多种数据源，如 DBMS、电子表格、平面文件。支持多种平台，支持多种数据库。具有规范的数据访问接口。具有灵活的可编程性和调用外部程序的功能。工具生成的代码必须是在开发环境中可维护的。能只抽取满足指定条件的数据和源数据的指定部分。能在抽取过程中进行数据类型转换和字符集转换并能计算生成衍生的字段。具有直观的视图、灵活的配置，能自动调用以定期实现管理工作。

# 3. 大数据时代数据融合与集成

## 3.1大数据的特点

在学术界广为人知的大数据“4V”特点表述是：“大数据”是以体量巨大（Volume）、类型繁多（Variety）、存取速度快（Velocity）、价值密度低（Value）为基本特征的数据集。大数据并不直接意味大价值，实际上是指经过分析发掘后可以释放潜在的价值。大数据还具有自身特点：

（1）多源异构的结构化数据和非结构化数据并存。

（2）大数据持续采样频率高，数据价值密度较低。

（3）数据蕴含信息复杂，关联性强。

（4）具有动态多时空时间序列性，数据不仅空间上具有相关性，时间尺度上数据也呈稀疏分布，且具有时间序列相关关系。

## 3.2大数据的来源

工业发展经历数字化装备和产品的普及、装备和产品网络化连接的普及（互联网+），进入网络化、智能化发展新阶段。制造业的数据早在20世纪70年代就可由计算机集散控制系统（Decentralized Control System，DCS）采集用于过程控制与设备状态监控的设备及传感器数据。随着企业工业化与信息化的融合应用，流程制造与离散制造企业不仅在时间上不断存储积累这些生产过程与设备运行数据，还在空间上扩展采集设备、人之间及内部传输的数据，从而获得时间与空间两个维度上不同尺度的大容量数据，以及分散于各生产部门的多源不同类型的文本、图像、声音等数据。工业大数据是以工业制造过程和工业产品为主体， 计算机工程与应用这些数据源于产品生命周期的各个环节，包括市场营销、设计、制造、服务、再利用等各个环节都会产生数据；企业外部市场、企业供应链的“跨界”数据也是工业大数据的重要来源。制造企业工业大数据具体包括。产品数据：设计 CAD 数据、产品建模仿真 CAE数据、工艺数据、加工数据、测试数据、维护数据、产品结构BOM 数据、零部件配置关系、变更记录等。企业运营数据：企业组织结构、市场营销与销售数据、业务管理数据、生产设备与产品本身的运行数据、质量检验与质量控制数据、生产制造数据、采购数据、库存数据、各类计划数据、电子商务数据等。价值链数据：客户、供应商、合作伙伴等销售、库存和计划等信息数据。外部数据：经济运行数据、行业销售数据、市场数据、竞争对手数据等。工业传感器、无线射频识别（RFID）、条形码、工业自动控制系统、企业资源计划（ERP）、计算机辅助设计（CAD/CAM）等技术更是日益丰富工业数据量。而工业企业中生产线高速运转，由机器所产生的数据量远大于计算机和人工产生的数据，而且数据类型多是非结构化数据。随着大规模定制和网络协同发展，制造业还需实时从网上接受众多消费者的个性化定制数据，并通过网络协同配置各方资源，组织生产，管理更多种类数据。

## 3.3大数据的融合和集成需求

在信息研究领域，信息抽取技术是一项必不可少的关键技术。面对如此海量的信息空间，如何更快更准确的抽取出用户感兴趣的内容是一个迫切需要解决的 问题，也是信息挖掘技术的一个重要研究方向。信息抽取不同于信息检索等信息 处理技术，它需要对文本进行命名实体的识别，并抽取出实体之间的关系，再加 上中文文本中词语的灵活多变、构词复杂且没有明显的标志，因此对中文命名实 体的识别及关系的抽取就显得更加困难。

目前，信息抽取的主要方法有两种，一种是基于知识库算法，这种方法需要建立一些规则，虽然这种方法的准确率较高，但是这种规则的确定是比较困难的，对编写者有较高的要求，且移植性不高；另一种是基于统计的机器学习算法，这种算法采用不同的模型，并利用人工标注的训练集进行学习，对于新的数据集则采用模型算出其相关的概率，并以此来得到最终的结果。这种方法代价较小，性能较高，便于移植，所以是当前研究的热点。

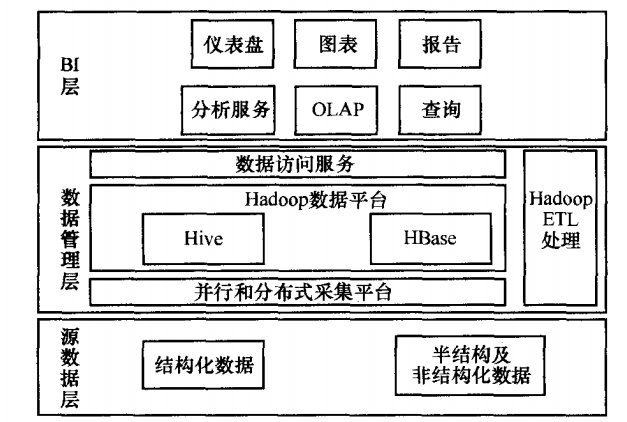
随着网络信息量的增大，对海量文本数据的信息抽取也变得复杂起来。如何利用海量文本数据来对实体关系进行更为准确的抽取是本文所要研究的一个重要问题。而这种大数据量的计算对算法的性能要求较高，采用何种策略来应对这种繁重的计算任务也是一个重要的问题。

大数据通常是由来源、主体或格式不同的数据合并而成，例如来自不同地区的调查数据，来自不同市场的金融数据，来自不同实验室的基因数据等。这种基于多个数据集的建模十分常见，了解不同子样本间的异质性(heterogeneity or difference)和同质性(homogeneity or similarity)是大数据分析的两个重要目标。但它的建模比较特殊，一方面，由于不同来源的数据存在差异，各不同数据源的同一变量的系数显著性和估计值可能存在差异，传统的处理方法是简单合并所有样本，建立统一模型，但是这种方法过于笼统，忽略了数据问的异质性(heterogeneity)；另一方面，也不能分开各自建立模型，因为这样会忽略各个数据集之间的关联性。数据集成分析方法同时兼顾这两方面，通过目标函数综合不同地区的数据，从统计角度考虑数据的异质性和同质性，以多个变量为研究目标，充分考虑了不同地区相互影响，同时求解多个模型。数据集成分析方法把不同来源、格式、特点性质的数据集中起来，相对于单一数据集模型，整合了更多的原始信息，能解决因地域、时间等因素造成的样本差异而引起的建模不稳定，在模型解释性和预测方面都具有显著优势。

传统数据仓库在这个数据爆炸的时代已经无法满足用户对数据应用的需求了，当然，传统数据仓库架构也仍在不断发展演化，这一点不容置疑，但也因其存在如上文所述的缺陷而在面对海量异构数据时表现的不够令人满意。现在有相当一部分的企业都拥有自己的数据仓库用来数据分析和数据挖掘，对于采用传统数据仓库的企业而言，大数据时代带来的机会就是使之能够利用过去无法通过传统数据仓库架构利用的数据，而Hadoop如今则为企业提供了运用新数据来源，使分析更加智慧的能力。因此，越来越多的企业开始构建自己的大数据平台来应对这个数据爆炸时代的数据应用需求，但是，同时也发现了一个问题：企业花费大量人力物力构建的传统数据仓库却成为了孤岛。于是，越来越多的企业开始考虑是否有一种通用模式可以将传统数据仓库和大数据平台整合在一起，建立一个统一的数据存储和数据处理架构，使得原来的传统数据仓库物尽所用，同时也能应对大数据的冲击?这成为一个急待解决的问题，该方案在已有传统数据仓库的基础上提供Hadoop的支持，弥补传统数据仓库在海量数据处理、存储等方面的不足，也可以依靠Hadoop的横向扩展能力突破单节点的传统数据仓库在存储和计算能力上的瓶颈，同时也重用企业费时费力构建好的传统数据仓库。

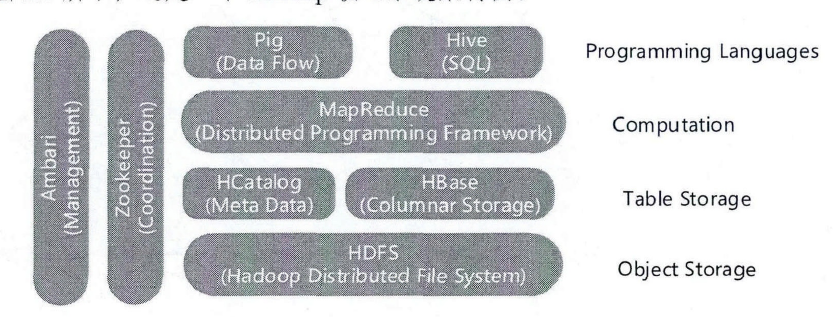
## 3.4大数据的融合与集成架构

从现有数据仓库建设理论和经验人手，引入部分大数据技术，特别是实现非结构化数据的收集、存储和处理是一种比较可行的方法。例如，将Hadoop技术应用于对数据的采集、ETL、存储、处理，开发提供给传统的数据仓库BI工具，其架构如图所示。



**图2**

在这个架构中，主要改变了传统数据仓库单节点数据处理和存储的方式，利用了Hadoop强大的数据处理能力，将各类数据处理成结构化数据，向上提供给传统BI工具，对数据进行分析和结果展示。在这个基础架构之上。可以 根据数据处理速度及分析响应能力，逐层进行细化及分解，优化组合 MPP数据库、内存数据库等各类技术，从而满足BI层分析展示的需要同。另外，还可以在数据管理层利用传统数据仓库和Hadoop共同合作，由传统数据仓库工具对结构化数据进行处理，由Hadoop对更大规模的非结构化数据进行预处理，并将两者处理后的数据存储至结构化数据库中，以便于BI层进行分析和展示。 除了技术层面上数据仓库与大数据的融合之外，非常重要的一点是传统数据仓库在具体应用理论、方法和实施上的成功经验，如基于数据驱动的螺旋式开发方法、调研及需求设计、ETL、数据建模、元数据管理唧等各个方面同样具有很多的可借鉴性与融合性。Hadoop核心展开图如下所示。



**图3**

其中HDFS为其提供分布式文件系统;Hcatalog可提供元数据存储及管理;HBase为一个基于列存储的分布式数据库;MapReduce为一个分布式数据处理模型及执行环境;Pig是用于检索大数据集的一种数据流语言和运行环境;Hive为一个提供类SQL查询语言的分布式数据仓库;Zookeeper则负责为庞大的分布式系统提供可靠的协调功能; Ambari为一种基于Web的工具,用于Hadoop集群的管理和监控。当然除此外,其主要项目还包括Mahout (提供可护展的机器学习和数据挖掘类库)、Avro(提供跨语言的远程过程调用以及数据的永久存储)、SqoopC为数据库和HDFS之间提供数据传输)等。接下来,我们将着重围绕HDFS、MapReduce、Zookeeper以及HBase等项目来阐述Hadoop平台。

HDFS分布式文件系统

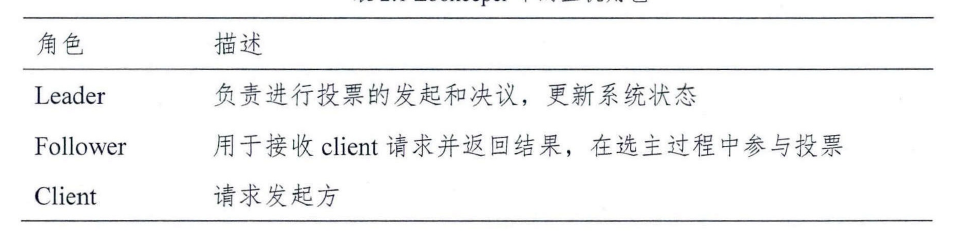
HDFS(Hadoop Distributed FileSystem)是Hadoop软件框架的分布式文件系统,为Hadoop项目的核心子项目。它的特性为:容错性(highly fault-tolerant)高,同时数据吞吐量(high throughput)大。这主要体现在:(1)硬件错误,即首先在一个含有大量服务器的HDFS系统中承认每个服务器均可能会出现硬件故障,继而能够检测到这种故障问题并采用内部机制保障其整体的可靠性;(2)提供流数据(Streaming Data Access)访问;(3)大文件支持,即HDFS能够对大文件(GB级别、TB级别甚至PB级别等)进行存储;(4)数据一致性保证,即HDFS支持“write-once-read-manyaccess"(一次写出、多次读取)的模型。

MapReduce分布式处理框架

MapReduce的核心内容主要由Map函数和Reduce函数构成,其采用分布式计算的方式来对海量数据进行处理并将最后计算所得结果存入指定的相应位置。简单而言,MapReduce就是“任务的分解与结果的汇总”。而用于执行MapReduce任务的机器角色有JobTracker和TaskTracker两种。前者主要在主节点上运行,后者则可运行在每个集群的从节点上。其中JobTracker用于调度Job,而TaskTracker主要用于执行Job。

Zookeeper分布式应用协同工具

Zookeeper可以为建立在Hadoop集群之上的分布式应用提供协调服务。其主要功能包括集群管理及分布式应用配置项的管理等。在Zookeeper中,主要存在三类机器角色，如下表所示



**图4**

在介绍HBase分布式数据库之前,我们不得不先提到NoSQL。这个概念最常见的解释“non-relational”,而“Not Only SQL”的说法也被很多人所接受。NoSQL是一种与RDBMS(关系型数据库管理系统)截然不同的数据库管理系统,它的数据存储格式可以是松散的、通常不支持Table之间的Join操作但能很好的支持横向I"展。于是我们不禁发问,为什么要使用NoSQL这是因为在大数据背景下,传统的RDBMS开始暴露出很多难以克服的问题,诸如:面对每秒上万次的写操作,磁盘的I/O巳无法承受;面对海量数据,存储记录数量有限且数据访问效率极低;横向扩展艰难,无法通过快速增加服务节点实现等。RDBMS的这种瓶颈催生了NoSQL的兴起及广泛应用。而HBase是一种基于列存储的NoSQL数据库,其为Hadoop中的一个子项目[26],依托HDFS作为其物理存储实现,同时也可通过MapReduce框架对HBase进行操作。

## 3.5比较传统数据仓库和大数据集成的特点

数据仓库是在企业管理和决策中面向主题的、集成的、与时间相关的、不可修改的数据集合，与其它数据库应用不同的是，数据仓库更像一种过程，对分布在企业内部各处的业务数据的整合、加工和分析的过程。建设数据仓库的最终目的是为企业提供决策支持。传统关系型数据仓库存在的一些不足，最为明显的有：一、现今企业传统数据仓库和关系型数据库擅长处理结构化数据，并且可以存储大量的数据，但传统数据仓库通常采用小机加盘阵高性能一体机建设，成本非常高，很大程度上限制了其扩展能力，即遇到性能瓶颈时，存储能力与计算能力的扩展是个很大的难题；二、传统数据仓库“擅长处理结构化数据”限制了可处理的数据种类，同时这种缺点还影响到传统数据仓库在面对海量异构数据时的表现，这通常意味着不少有价值的数据源是传统数据仓库无法涉及的。

大数据集成平台的有点有这些：1、支持超大文件。一般来说，HDFS存储的文件可以支持TB和PB级别的数据。 2、检测和快速应对硬件故障。在集群环境中，硬件故障是常见性问题。因为有上千台服务器连在一起，故障率高，因此故障检测和自动恢复hdfs文件系统的一个设计目标。假设某一个datanode节点挂掉之后，因为数据备份，还可以从其他节点里找到。namenode通过心跳机制来检测datanode是否还存在 3、流式数据访问。HDFS的数据处理规模比较大，应用一次需要大量的数据，同时这些应用一般都是批量处理，而不是用户交互式处理，应用程序能以流的形式访问数据库。主要的是数据的吞吐量，而不是访问速度。访问速度最终是要受制于网络和磁盘的速度，机器节点再多，也不能突破物理的局限，HDFS不适合于低延迟的数据访问，HDFS的是高吞吐量。 4、简化的一致性模型。对于外部使用用户，不需要了解底层细节，比如文件的切块，文件的存储，节点的管理。因为存储在HDFS上的文件都是超大文件，当上传完这个文件到hadoop集群后，会进行文件切块，分发，复制等操作。5、高容错性。数据自动保存多个副本，副本丢失后自动恢复。可构建在廉价机上，实现线性（横向）扩展，当集群增加新节点之后，namenode也可以感知，将数据分发和备份到相应的节点上。 6、商用硬件。Hadoop并不需要运行在昂贵且高可靠的硬件上，它是设计运行在商用硬件的集群上的，因此至少对于庞大的集群来说，节点故障的几率还是非常高的。HDFS遇到上述故障时，被设计成能够继续运行且不让用户察觉到明显的中断。

大数据需求的产生背景与数据仓库类似，人们希望利用新技术处理越来越多的数据、挖掘更大的数据价值。因此从需求角度来说．无论是数据库、数据仓库还是大数据都是解决不同需求、处理不同级别数据量的技术，它们之间并无冲突．所以短期内并不会出现由谁取代谁的结果，而应该是针对不同需求和现状进行技术选择，各种技术相互补充、相互协作。

# 4. 数据融合与集成趋势

新时期下，一方面，以大数据为基础的人工智能的飞快发展，对数据融合与集成有新的要求。人工智能时代下，各种各样的数据都可以通过传感器进行采集的，这意味着各种传感器融合成为当下数据融合与集成的关键。另一方面，大数据时代下，更多的半结构化和非结构化的数据，而这些数据的融合与集成是数据融合与集成的难点。下文主要是从这两方面展开。

## 4.1人工智能下传感器数据集成与融合

大数据是人工智能的基础，在人工智能时代下，大量的数据是通过智能的传感器终端代替人工输入来获得的。比如在自动驾驶中，路面的状况，周边的环境等都是通过相关的传感器进行获取的，然后把获取到的数据进行融合与集成，再传输到驾驶大脑进行处理，最终达到自动驾驶汽车。传感器的数据的集成与融合在人工智能中扮演着重要的角色。

广义地说,传感器的融合就是多传感器的集成化使用和设计,常称为多传感器集成(MSI)。这种集成比分离使用的传感器组合有更好的性能。多传感器集成(融合)的优点是:

* 成本低,质量轻,占空间少
* 增加测量维数,增加了置信度
* 容错性好,性能稳定
* 改进探测性能,增加响应的有效性
* 扩展了空间和时间的覆盖
* 改进系统的可靠性和可维护性
* 提高了空间分辨率
* 降低了对单个传感器的性能要求

多传感器的融合要考虑传感器性能的互补性,电磁兼容性及资源共享等问题。多传感器的集成度和相关的处理功能对多传感器系统的设计是一项重要的考虑。下面是4种传感器集成度的例子及特点:

1. 0集成度:例如制导雷达,夜视仪等。其特点是:
2. 分离的传感器“捆在一起”,置于一个平台上
3. 每个传感器的功能是独立的
4. 从一个传感器来的数据使用时与其它传感器的数据无关
5. 1集成度,信号集成:例如F-14的雷达和跟踪控制系统,其特点是:
6. 分离的传感器“拴在”一个平台上
7. 从一个传感器来的数据同其它传感器来的数据组合,或用于其它传感器的工作控制
8. 2集成度,物理集成:例如光电火控系统,TADS,阿伯奇,F-16等,物理集成+AI可到更精确的融合，特点是:
9. 各传感器作为一个整体设计
10. 各传感器既可以物理上配置在一起,也可以提供传感器相对方向的配置
11. 在可能的地方孔径共享
12. 从一个传感器来的数据与从其它传感器来的数据组合,或控制其它传感器的工作
13. 集成度,融合集成:例如人类智能，特点是:
14. 一个传感器数据的分析受到来自其它传感器数据的影响,反之亦然
15. 数据处理呈现:整体大于各部分之和
16. 物理集成是基础

### 4.1.1多传感器数据融合面对的问题

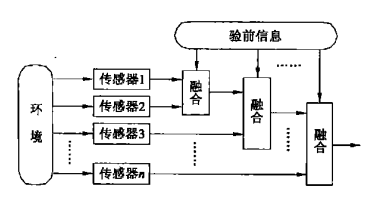
在多传感器系统中,每个传感器得到的信息都是某个环境特征在该传感器空间中的描述。由于各传感器物理特性以及空间位置上的差异,造成这些 信息的描述空间各不相同,因此很难对这样的信息 进行融合处理。为了保证融合处理的顺利进行,必须在融合前对这些信息进行适当的处理 ,将这些信息映射到一个共同的参考描述空间中,然后进行融 合处理 ,最后得到环境特征在该空间上的一致描述。融合处理的前提条件是从每个传感器得到的 信息必须是对同一目标的同一时刻的描述。这包括 两个方面,首先要保证每个传感器得到的信息是对 同一目标的描述,比如同一物体的位置信息。在多传感器融合中,这被称之为数据关联。其次,要保证各传感器之间应该在时间上同步。在动态工作环境下,同步问题表现得尤为突出。

### 4.1.2传感器资源信息元数据模型

元数据是关于数据的数据, 它为各种形态的信息资源提供规范"普遍的描述方法和检索方式!用于描述数据的内容"质量"管理方式"数据的所有者"覆盖范围"数据的提供方式等信息.元数据是分布式的"多源的信息资源实现共享"整合以及互操作的工具与纽带!是数据与数据用户之间的桥梁.要实现万维网环境下城市异构传感器资源标准地共享,就要确立传感器资源共享元数据,它应包含传感器资源的标识"分类"发现"访问控制等信息.

传感器资源共享八元组元数据(metadata,MD):

MDi={MDiG,MDiCS, MDiPP, MDiCT, MDiA, MDiGP, MDiI, MDiPC}(I=1,2…,N),其中, MDiG代表通用信息元数据; MDiCS 代表约束元数据; MDiPP代表属性元数据; MDiCT代表联系元数据; MDiA代表存档元数据; MDiGP代表地理定位元数据; MDiI代表接口元数据; MDiPC代表过程元数据. 进一步对前面的传感器信息按属性归类于元数据类型:传感器分类与标识信息主要为通用型元数据;传感器观测有效时间,传感器共享级别与合法性等属于约束型元数据)传感器固有特征以及观测,通信,计算能力等为属性元数据)传感器负责单位或个人,传感器在线引用等信息为联系型元数据;传感器或传感器数据服务发布时间,传感器在线文档链接等信息为存档型元数据!传感器及其搭载平台所在时空坐标系;传感器观测系统的动态或静态空间观测位置信息为地理定位元数据"传感器可得性服务,如传感器规划服务与传感器观测服务为接口型元数据,传感器观测数据所涉及到的处理,包括输入,输出,参数以及处理方法信息为过程型元数据.



**图5**

### 4.1.3多传感器融合方法

作为多传感器融合的研究热点之一 ,融合方法 一直受到人们的重视 ,这方面国外已经作了大量的 研究 ,并且提出了许多融合方法 [12～ 14 ]。 目前 ,这些 方法大致可分为两大类: 概率统计方法和人工智能 方法。 其中人工智能方法又可以分为两种: 逻辑推 理方法和学习方法。

概率论已有很长的历史 ,它成功地处理了许多与不确定性有关的问题 ,有丰富的理论和系统的方法。它所研究的现象是随机的,用随机变量来表示 不确定信息 (随机信息) ,将概率统计方法作为不确 定信息处理的手段。在多传感器融合中 ,常采用的 与概率统计有关的方法包括: 估计理论[15]、卡尔曼 滤波[16]、假设检验[ 17, 18 ] 、贝叶斯方法 ,统计决策理 论 [19]以及其他变形的方法 [20, 21 ]。其中,文 [12]对卡尔曼滤波、贝叶斯估计、多贝叶斯方法和统计决策 理论做了简单的介绍 ,其他方法可以在本文所列举 的参考文献中找到。概率统计方法可以在融合的各个层次上使用。尽管概率统计方法已经成为多传感 器融合研究中不可缺少的工具,但是在不确定性推 理中仍然存在一些问题。 这主要有两方面的原因,首先,概率论本身存在着缺陷。目前,对概率存在两 种不同的解释,一种是客观概率,即事件发生的频 率;另一种是主观概率,即人的信念。这两者既存在着联系,又存在着差异和矛盾。此外,概率论在传统 框架下无法综合不同信息所表示信念。在文[22]中,张尧庭等人讨论了概率论存在的问题,并且介绍了两种改进的概率方法,即似然比推理和信任函 数理论。这两种方法可以在一定程度上克服传统概率方法所存在的问题。其次 ,在许多融合方法中,往 往对传感信息的性质做了一些明确的假设,最常见 的假设就是每个传感器所采用的测量模型都包含一附加的白噪声项,同时假设各个传感器之间相互独立。可是在实际应用中这一假设很难满足,这对 概率统计方法的实际应用会造成影响。此外,由于许多概率统计方法都是基于一个确定的概率分布,因此大多数概率统计方法只适合于静态的工作环境。

逻辑推理包括概率推理、证据推理、模糊推理、和产生式规则等。 它们都属于不确定性推理,是人工智能研究中最为活跃的研究领域之一。不确定性推理的方法 ,首先需要对不确定信息进行表示(或度量)、不同的表示方法即构成不同的不确定性推 理的方法。文 [23 ]对人工智能中的不确定性进行了深入的讨论。文 [24]对不确定性推理原理进行了详细的讨论,并且提出了包含度理论。包含度理论不仅是研究不确定性推理的一般原理 ,而且是研究不确定现象的方法学。文 [25, 26]分别给出了模糊方法和D-S证据理论在多传感器数据融合中的应用。文[10]介绍了产生式规则在多传感器融合中的应用。此外, Luo和 Kay还对 D-S证据理论、模糊 逻辑和产生式规则进行了比较[13],与概率统计方法相比,逻辑推理存在许多优点,它在一定程度上克服了概率论所面临的问题 ,它对信息的表示和处理更加接近人类的思维方式,它一般比较适合于在 高层次上的应用 (如决策),但是逻辑推理本身还不74数据采集与处理第 19卷够成熟和系统化。此外,由于逻辑推理对信息的描述存在很大的主观因素,所以信息的表示和处理缺 乏客观性。和概率统计方法一样,逻辑推理在实际应用中也存在一些问题。例如,证据组合规则是DS证据理论的核心,但在应用中要求满足组合证据 之间相互独立,这在实际应用中往往难以满足在文[27]中,针对这一问题对D-S证据理论做了一些改进。

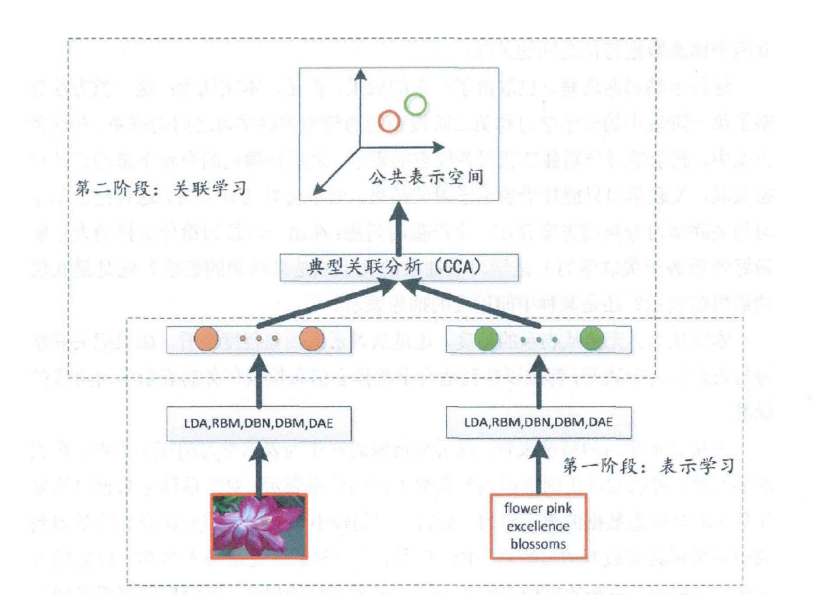
学习方法在多传感器融合的研究中还没有引 起人们的足够重视。这主要是因为学习方法在理论 方面还不够完善 ,需要更加深入的研究。目前多传感器融合的研究主要集中在工作环境为静态的情 况下 ,很少考虑未知与动态的情况。在这方面,Luo 和 Kay曾多次强调适用于动态与未知情况的融合 方法将会成为多传感器融合未来研究方向之一,同时还强调人工智能领域,尤其是神经网络将会为多传感器融合的研究在理论和实践上提供更加有力 的手段 [12, 13 ]。通过对这些方法的研究,不难发现用学习方法来解决多传感器融合问题有以下几个特点:首先,能够有效地解决不确定、动态情况下多传感器融合存在的一些问题,如复杂操作的建模问题,学习问题,鲁棒性问题等;其次,多采用学习算 法作为实现的主要手段,通过训练来得到复杂操作的模型;最后,由于学习算法自身具有处理不确定 性的能力 ,因此不需要其他复杂的不确定信息处理 方法。因为这些特点,使得系统具有良好的性能 ,如自适应性和鲁棒性。目前 ,学习方法自身还存在一些问题,例如,稳定性问题、泛化能力、缺乏有效的学习机制等。目前已有的学习方法包括神经网络、映射学习方法、数据挖掘等[11, 28～30 ]。作者目前也从事这方面的研究工作,研究的内容主要是能够适应环境动态变化的 具有学习能力的融合方法。尽管对目前已有的融合方法进行了分类介绍,但它们之间没有严格的界限。例如,在人工智能中,利用概率主观的一面产生了基于概率的不确定性推理。此外,各种方法之间又存在一定的联系,而且它们之间的组合可以形成性能更好的融合方法。文[31]利用产生式规则和神经网络来实现可选传感 器融合系统。文[32]用神经网络和贝叶斯法则得到一种自适应传感器模型。每种融合方法都存在各自的优缺点,显然这种组合可以更好地发挥它们各自的优点。

## 4.2大数据下非结构化(多模态)数据的融合

大数据时代下，数据不仅有结构化的，更有半结构化和非结构化的数据。结构化的数据往往存放在关系数据库，非关系数据库中。而半结构化和非结构化的数据往往是以文本，图片，视频等形式存在。大数据时代下，半结构化和非结构化的数据是海量的，而数据仓库和基于Hadoop的大数据融合架构更多的是处理结构化的数据。下文将会介绍非结构化数据的融合。

非结构化的数据，一般以文本，图片，音频，视频等形式存在。这些不同模态的非结构化数据往往是以不同模态的数据来描述统一个事物。比如，中国的首都北京，可以通过文本，图片，视频等来描述。正是因为不同模态的数据描述了同一个事物，可以通过描述同一个事物的不同模态的数据的融合，从而达到非结构化数据的融合。

为了实现大数据下的多模态数据融合，各领域出现了一些融合方法，可以认为普遍采用了如下图所示的方式，如图1所示：



**图6**

当下这种模式的所需要的关键技术有特征提取，隐藏层特征提取，隐藏层特征关联，构建隐藏层特征关联图。其中，1）特征提取用于提取各模态异构数据的特征；2）隐藏层特征提取是运用深度网络模型对特征进行学习与提取；3）隐藏层特征关联是多模态数据融合的关键一步；4）构建隐藏层特征关联图,是多模态数据融合的形象体现，可以用于对多模态数据的检索。

### 4.2.1特征提取

文章对于多模态数据的形式主要是以图片和文本这两种模态。对于原始的图像和文本数据，我们需要对它们进行特征提取，把原始的图像和文本通过向量的形式表现出来，这样就可以用机器学习的方法对他们进行处理。

**1 文本特征提取**

文本特征提取是从文本中抽取出特征词进行量化表示文本信息，它将一个无结构的原始文本转化为结构化的计算机可以识别处理的信息，即对文本进行科学的抽象，建立数学模型，用于描述和代替文本。目前，特征选取有4种方式：（1）用映射或变换的方法把原始特征变换为较少的新特征；（2）从原始特征中挑选出一些最具代表性的特征；（3）根据专家的知识挑选最有影响的特征；（4）用数学的方法进行选取，找出最具有分类信息的特征。具有这4种方式产生了各种各样的算法与模型，比如词袋模型，TFIDF，词频方法，互信息，二次信息熵等等。文章将会详细介绍词袋模型。

词袋模型是用映射或变换的方法把原始特征变换为较少的新特征这种方式简化地对文本进行表示。这种模型通过一袋单词来表示一篇文档，忽略了语法以及单词在文本中出现的次序，也就是说文本中每个词的出现都是独立于其他单词的。词袋模型把一篇文章示成了文章里每个单词出现的频度的向量，向量中的每个元素表示词典中该索引号所对应的单词在文章里出现的频度。比如说由如下两篇文档：

I like to play football. Tom likes play football too.

James likes to play football too.

基于上述两篇文档，首先构建单词表：

{“I”:1,”like”:2,”to”:3,”play”:4,”football”:5,”Tom”:6,”likes”:7,”too”:8,”James”:9}

该单词表一共包含 9个单词，当然在实际运用中会根据停用词表去掉停用词，

基于该单词表上述两篇文档的词袋模型表示为：{1,1,1,2,2,1,1,1,0},{0,0,1,1,1,0,0,0,1}。每一个向量中的元素代表了字典中与之相对应序号的单词的出现次数。比如说，第一个向量的第四个元素为 2，就表示词典中第二个单词”play”出现的次数为 2。

如上文所述，很多的文本长短有差异，如果采用所有单词来构成字典会使词袋模型向量维度过大，导致词袋模型向量过于稀疏，从而影响效果，所以必须去除出现频度过少的单词。此外，对于一些无关的停用词，比如像“我”、“你”等人称代词，还有一些介词等等，所以我还必须构建停用词表，并在构建词典前还需要对文本进行过滤。具体的单词过滤步

骤如下：

1. 去除停用词
2. 去除长度过短的单词
3. 去除频度过少的单词

上面主要是介绍词袋模型的表示方法和单词过滤。一般的文本特征提取的步骤如下：

1. 输入所有的文本
2. 对每篇文本进行过滤，去除其中的停用词等。
3. 对所有的文本进行遍历，提取单词，记录每篇文档中出现的单词及词频。
4. 利用每篇文档的单词及词频构造单词表。
5. 对所有文本再进行一次遍历，根据词典来构建词袋模型向量。
6. 输出每篇文本的特征矩阵。

**2图像特征提取**

图像的特征提取是使用计算机提取图像信息，决定每个图像的点是否属于一个图像的特征，提取的结果是把图像上的点分为不同的子集。由于实际问题的应用类型各有不同，采取的方法也是各式各样的，一直以来图像特征提取都没有一个精确的定义。比较常见的图象提取方法有颜色特征提取，纹理特征提取，形状特征提取，空间关系特征提取。本文主要是详细介绍居于形状特征提取视觉单词词袋模型（BOVW）。在视觉单词词袋模型中，我们可以将图像看作是一篇文档，把视觉单词看作是普通文本单词。由于图像中的视觉单词不像文本有现成的，需要采用基于SIFT算法的形状特征提取对图像进行特征提取。

为了把图像用视觉单词词袋模型表示出来，必须进行三个步骤：提取视觉单词，构建词典，图像的词袋模型表示。

提取视觉单词

视觉单词的提取一般采用SIFT算法。SIFT(尺度不变特征变换)是由David G.Lowe提出的一种基于尺度空间的，对图像缩放，旋转甚至仿射变换保持不变性的图像局部特征描述算子。该方法将一幅图像映射为一个局部调特征向量集，这些特征集具有平移，缩放，旋转不变性，并且对光照变换，仿射及投影变换也有一定不变性。可以采用开源库OPENCV提取图像的SIFT特征，通过SIFT特征提取后，一幅图像可以表示成一系列SIFT特征描述符的集合，每一个SIFT描符是一个128维向量。其过程如下：

1. 构建尺度空间，利用高斯卷积核实现多尺度变换。生成尺度空间。
2. 检测尺度空间极值点
3. 精确定位极值点
4. 为每个特征点确定方向参数
5. 生成图像集SIFT描述符

**3 构建词典**

经过SIFT特征提取后，一个SIFT描述符可以描述一个视觉单词，但由于一个SIFT描述符是一个128维向量，一般的视觉单词是一义多词的，所以需要对SIFT描述符进行聚类，使用统一的一个视觉单词来表达一个含义。聚类方法有K-means聚类算法，层级聚类算法，SOM聚类算法，FCM聚类算法。本文详细介绍K-means聚类算法。K-means算法是一种基于样本间相似度量的间接聚类方法，属于非监督学习方法。K-means算法把n个对象分成k个族，族内具有较高的相似度，族间相似度较低。对SIFT描述符集合聚类，假设有图像数据集 T={ t1,…tn}， i t 是一个图像的 SIFT 描述符集合，不同图像所包含的 SIFT 描述符数量不同，k个单词的词典，其过程如下：

1.随机选取 K 个聚类质心，分别为{u1,…uk}每个质心是一个 128 维的向量。

2.对于图像数据集中的每一个图像的 SIFT 描述符集合 t i={ x1,…,xm}, xi 是一个128 维 的 SIFT 描 述 符 ， 计 算 向 量 中 的 每 个 点 i x 应 该 属 于 哪 个 簇 ：ci=argmin||xi-uj||2,其中

uj=

3.重复过程 2 直到收敛为止其中 i c 代表样本点 i x 与 K 个单词中距离最近的那个单词的序号，即 c i€{1,…,K}，质心 j u 代表序号为 j 的样本中心点向量，具体步骤如下：首先随机选取 K 个质心，接着计算图中每个样本点到 K 个质心的距离，然后选取距离近的那个样本点簇，重新计算每一个样本点簇的质心(对簇内所有样本点坐标求均值)，重复迭代上述步骤直到质心变化很小。经过以上步骤包含 K 个单词的词典已经建立好。

**4.图像的词袋模型表示**

根据图像词典，对SIFT特征描述符进行映射，求取离每个SIFT特征描述符最近的单词，用该单词的序号来代替原始的SIFTt特征描述符，然后统计词典中的每个单词在图像特征向量中出现的次数，就可以把一幅图像表示成视觉单词词袋模型的形式。

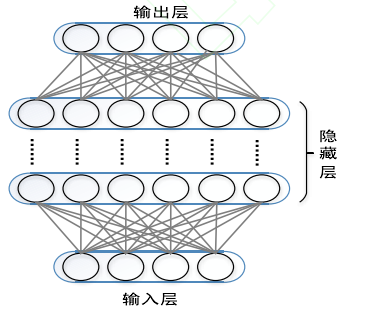
### 4.2.2深度网络隐藏层特征提取

对于原始的图像和文本数据进行预处理和特征提取，把它们用词袋模型的形式表达出来。这样的表达，只是对原始文本和图像的向量表示，它们两者差异巨大，缺乏关联性。同时这些特征是高纬度，高稀疏的。这样会大大影响融合的效果和后续的数据挖掘。通过深度网络模型，实现原始特征向量到隐藏层特征向量的映射，获得原始特征的高层抽象，经映射后也降低向量的维度。

**1.深度网络模型**

深度学习(Deep Learning)作为人工智能技术的展开研究中一种新的机器学习方法，具有非常重要的意义。它以云计算对大数据的并行处理能力和算法为基础，构建一个更加贴近于人脑思维方式的学习网络，使计算机找到如何处理“抽象概念”的方法，使计算机更加智能。深度学习网络的学习算法属于无监督特征学习，它以特征层次结构学习为思想基础，通过对底层结构特征的组合与学习形成更抽象的高级特征，用更为抽象的高级特征代替原始特征表达事物的属性。深度网络拟在模拟人脑的思维机制来解释数据，构建具有分析和学习能力的深度神经网络，模拟人脑的思维方式解释和学习数据，以发现数据的特征。

神经网络技术起源于上世纪五、六十年代，当时叫感知机，是最早被设计并实现的人工神经网络，是一种二分类的线性分类模型，主要用于线性分类且分类能力十分有限。输入的特征向量通过隐含层变换达到输出层，在输出层得到分类结果。早期感知机的推动者是 Rosenblatt。但是单层感知机遇到一个严重的问题，即它对稍复杂一些的函数都无能为力（比如最为典型的“异或”操作），随着数学理论的发展，这个缺点直到 20世纪 80 年代才被 Rumelhart、Williams、Hinton、LeCun 等人发明的多层感知机（multilayer perceptron，MLP）克服。多层感知机可以摆脱早期离散传输函数的束缚，使用 sigmoid 或 tanh等连续函数模拟神经元对激励的响应，在训练算法上则使用Werbos 发明的反向传播 BP 算法。



**图7 全连接深度神经网络结构示意图**

图7为全连接深度神经网络结构示意图。通过增加隐含层的数量及相应的节点数，可以形成深度神经网络，深度神经网络一般指全连接的神经网络。

神经网络主要有两种体系结构：前馈网络和反馈网络。前馈网络只允许信号在一个方向上传递，即从输入层向输出层传递。如图 2-6 所示，前馈网络没有反向回馈，任何层的输出不能影响同一层及前面的所有层。前馈网络倾向于构建一个前向无回路的网络。反馈神经网络通过循环让信号在两个方向上传播。反馈网络的结构所示，反馈神经网络的结构异常复杂，反馈神经网络具有动态性，他们的状态在不断变化，直到达到平衡点为止。他们保持一个平衡状态直到输入变化或者需要找到一个新的平衡。

**2.隐藏层特征提取**

借鉴深度学习的方法，通过深度网络模型把文本和图像的的原始特征映射到更抽象的隐藏层特征，使得到的隐藏层特征具有相同的维度。可以通过分别学习两个L层神经网络模型把高维度的原始特征向量变成低维度的隐藏层特征向量。

文章采用两种模态的数据(图像和文本 ) X=[x1,..xn],€Rdx\*nx,Y=[y1,..yn]€Rdy\*ny,X 和 Y 分别是图像数据集跟文本数据集的特征矩阵，dx 和dy 分别表示两种模态特征的维度，由于已经对底层特征进行了预处理，采用了词袋模型的表示形式，所以dx 和dy 即分别表示图像跟文本的词典所包含的单词数， x n 和 y n 分别表示数据集 X 和 Y 中的样本数量。我们用 n 来表示训练集的样本数，用 m 来表示测试集的样本数，我们用训练集来构建模型。此外，T=[t1,…tn]€Rc\*n是训练集的类别标签矩阵，每一个ti€Rc是第 i 对训练集的类别标签向量，c 是类别标签总数。 tij=1 表示第 i 对训练集的类别标签为第 j 类，其他都为 0。所有类别标签向量都采用归一化处理，也就是说某一对训练集的所有类别标签总和为 1。神经网络模型一般分为训练和测试两个阶段。我们需要使用训练集的样本来构建深度网络模型。由于两个神经网络模型构建的原理和过程基本一致的。此处选取的文本原始特征来叙述模型的构建过程，图像原始特征模型的构建也亦然。深度网络模型构建步骤如下：

1.初始化神经网络模型的每一层

2.选取训练集中的每一个向量 xi ,以及与之相对应的类别向量 ti ,通过获得的误差函数，将误差反向传播，按梯度下降的方法微调各层的权值，不断循环该过程直到误差达到目标值为止，保存所获得模型

3. 返回深度网络模型.

运用构建好的深度网络模型，输入测试集的样本数据，我们就可以获取隐藏层特征。

### 4.2.3隐藏层特征融合

获取的文本以及图像的隐藏层特征，由于分别使用了两个深度网络模型来获取图像以及文本的隐藏层特征，不同模态的隐藏层特征并没有进行关联，他们缺乏相关性。本文主要详细介绍基于典型相关性分析把文本与图像的隐藏层特征向量映射到一个同构的媒介空间，使得映射后的隐藏层特征向量不仅具有同维度，且具有最大的相关性，从而达到隐藏层特征融合。

CCA典型相关分析（canonical correlation analysis）利用综合变量对之间的相关关系来反映两组指标之间的整体相关性的多元统计分析方法。它的基本原理是：为了从总体上把握两组指标之间的相关关系，分别在两组变量中提取有代表性的两个综合变量U1和V1（分别为两个变量组中各变量的线性组合），利用这两个综合变量之间的相关关系来反映两组指标之间的整体相关性。

1936年，Hotelling提出典型相关分析。考虑两组变量的线性组合, 并研究它们之间的相关系数p(u,v).在所有的线性组合中, 找一对相关系数最大的线性组合, 用这个组合的单相关系数来表示两组变量的相关性, 叫做两组变量的典型相关系数, 而这两个线性组合叫做一对典型变量。在两组多变量的情形下, 需要用若干对典型变量才能完全反映出它们之间的相关性。下一步, 再在两组变量的与u1,v1不相关的线性组合中, 找一对相关系数最大的线性组合, 它就是第二对典型变量, 而且p(u2,v2)就是第二个典型相关系数。这样下去, 可以得到若干对典型变量, 从而提取出两组变量间的全部信息。典型相关分析的实质就是在两组随机变量中选取若干个有代表性的综合指标（变量的线性组合）, 用这些指标的相关关系来表示原来的两组变量的相关关系。

tr\_X[n][ Ax]和 tr\_Y[n][ Ay]为通过深度网络训练后保存的图像训练集和文本训练集的隐藏层特征矩阵。为了方便起见我们用 X [x1,...xn]来表示 tr\_X[n][ Ax]，其中Ai €RAx表示图像训练集第 i 个样本的隐藏层特征向量；同样的我们用Y [ y1,...yn]  来表示 tr\_Y[n][ Ay]，其中 Ai € RAy表示文本训练集第i个样本的隐藏层特征向量。定义两个矩阵 A 和 B 分别为两个系数矩阵，分别用来对 X 和 Y 进行线性组合，则可以把我们所要求的基向量组成的矩阵表示为 U= ATX，V= B TY ,使用皮尔森(Person)相关系数来定义 U 跟 V 的相性,由此问题变成了获取一组最优解 A 和B 使得相关系数最大。经过一系列变换变化后要解决典型的特征根问题即可求解系数矩阵 A 和B。求得系数矩阵 A 和 B，我们就可以用系数矩阵对图像和文本的隐藏层特征向量进行投影，使经过投影后的向量具有相近意义表达。

# 5. 案例分析

## 5.1智慧城市的数据融合与集成应用

智慧城市是近年的研究热点，它是以信息技术为支撑，具备健全，透明，充分的信息获取；通畅，广泛，安全的信息共享；有效，规范，科学的信息利用。从而使得[3]：政府具有更加高效的决策和执行力；企业能够更加高效和低成本的运行；大学可以更加精准和个性化地培养人才；百姓能够获得更加便捷和舒适的生活；同时，能够催生智慧校园，智慧交通等一大批新的产业链条，产业圈。智慧城市就是要借助多种设备感知城市中的各种数据（多模式数据），形成多模式数据感知系统，并进行城市数据的融合，为城市的各大主题提供智慧的服务[4]。因此智慧城市的建设得到了各界的广泛研究和政府的大力支持。

智慧城市建设的基础就是对城市多模式数据的感知和挖掘。这些数据是城市的战略资源[5]，通过对这些数据进行挖掘、融合和充分利用，其影响力远远超出人们的想象。数据挖掘能够发掘无数有趣的或者与生活有关的事情。

下面以数字冰雹公司为例[6]：

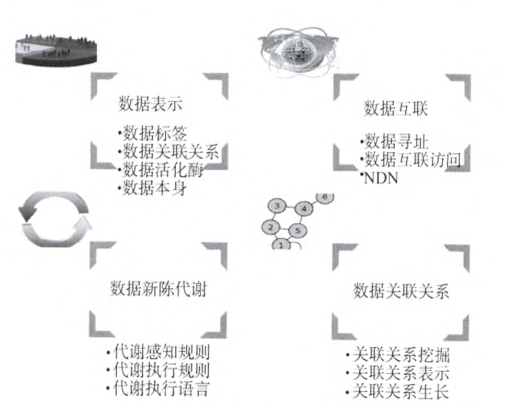


**图8**

智慧城市大数据可视化决策平台主要为政府部门提供城市运行监控，产品基于城市正常运行态势，集成地理信息、GPS数据、建筑物三维数据、统计数据、摄像头采集画面等多类数据的可视化决策平台，可以把政府各个部门、各种类型的数据融合打通，包括市政、警务、消防、交通、通讯、商业等汇集在统一的平台之上。

获取这些海量的数据之后，为了实现数据互联融合，数据的表示也需进行大的改变。不再是只表示数据本身，更重要的是在融合场景下，数据包括的内容将更加丰富，需要能对数据所包括的所以内容都能进行表示。包括数据关联关系的表示，数据虚拟标签的表示。同时，需要描述数据活化酶，它像生物体中的酶，类似一个数据处理者的身份，可以促进数据融合的自动进行。

然后进行数据互联，数据互联的目的是建立数据之间彼此沟通的通路，形成数据网络，也是融合的一部分。为了实现这一目标，从互联的角度来看，除了数据虚拟标签能实现数据的唯一标识外，还需要解决数据寻址、数据互联访问题。



**图9.数据融合互联模型**

经过上述的操作，我们将数据融合起来分析，反馈到平台，进行可视化的操作，如下图所示：



**图10.** **可视化情报分析平台**

该平台针对多元异构的海量数据，通过数据处理、存储管理、可视化交互分析等技术，实现图形化情报信息查询、可视化关联分析、证据链和情报线索发掘等功能。平台旨在利用人判断，把人的思想和计算机的运算能力结合，通过直观的可视交互手段，帮助用户捕捉到对象之间的关联关系。

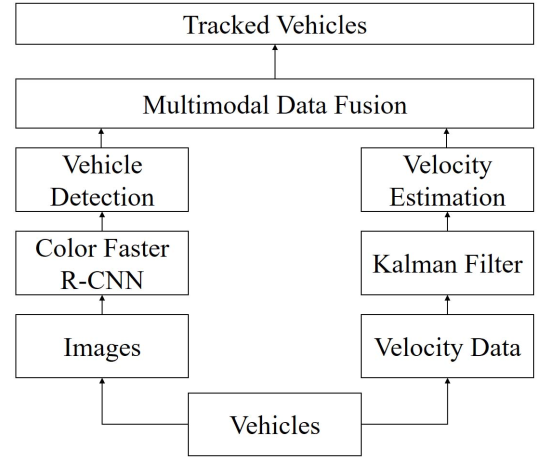
案例分析总结：

上述中例子是大数据下的智慧城市，是近几年研究的热点，在构建智慧城市的过程中，数据源的种类多样[7]，数量巨大。例如，在城市路网中，各个十字路口以及主要干道上都会出现监视器，摄像头。这些监视器与摄像头一方面能够快速捕捉城市地面上车辆的行车情况（如车速，车距）[8],另一方面也为针对城市数据挖掘提供了较好的素材。有了大数据之后，要能成功构造智慧城市，离不开多模态数据融合分析，数据融合分析是建成智慧城市必要条件。

## 5.2基于多模态数据融合的车辆跟踪应用

车辆位置预测或车辆跟踪是连接车辆中的重要主题。但是，如果只有单一的模态数据，这个任务是困难的，可能会造成偏见并阻碍准确性。随着连接车辆中的传感器网络的发展，多模式数据正变得易于获得。因此，我们提出了一个多模态数据融合的车辆跟踪框架。具体而言，我们将两种模态，图像和速度的结果融合在我们的车辆跟踪任务中。在车辆检测模块中处理的图像提供关于车辆特征的视觉信息，而速度估计可以进一步评估目标车辆的可能位置，这减少了被比较的候选者的数量，减少了时间消耗和计算成本。我们的车辆检测模型的设计采用了色彩更快的R-CNN，其输入特征是车辆的纹理和颜色。同时，速度估计是通过卡尔曼滤波器实现的，这是一种经典的跟踪方法。最后，应用多模式数据融合方法来整合这些结果，从而实现车辆跟踪任务。

下面是具体车辆跟踪的总框架：



**图11.** **多数据融合车辆跟踪的总框架**

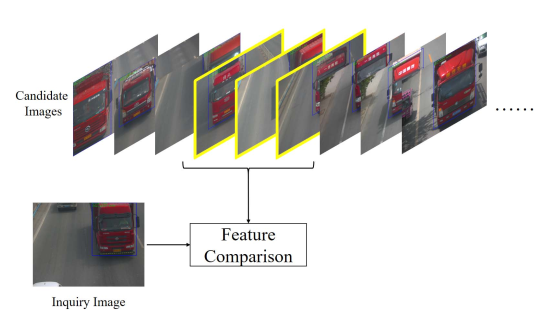
上述提出的多模态数据融合方法的体系结构。这个框架中存在两个并行模块。左边是用摄像机获取车辆的图片，用深度学习方法来获取车辆的纹理和颜色特征。右边车辆速度检测模块，是用来估计速度。主要基于运动传感器获得的数据进行卡尔曼滤波。通过上述的操作，我们就获得了车辆图片数据和车辆速度数据。最后，通过多模态数据融合方法融合了这两种方法。实现车辆跟踪任务。

车辆跟踪的融合过程：我们获得车辆图片后，用上述的深度学习方法，输入车辆的纹理和颜色特征，在跟踪过程中，我们得到了一系列候选图片，这些候

图片数量是巨大的，都来自大小为M的矩阵T。因此，在另外一个数据，车辆速度数据就起到了过滤的作用。其中，车辆速度数据F是一个N1的向量。且F符合高斯正态分布。将两者的数据进行融合计算：

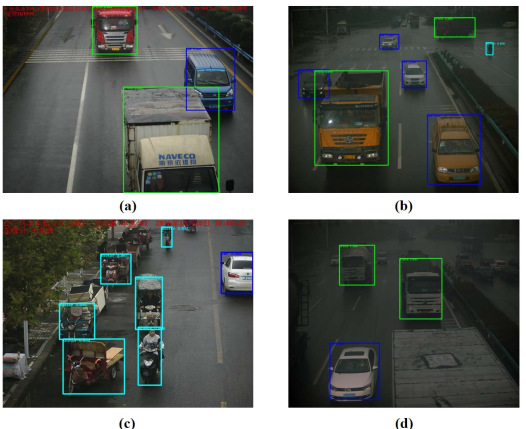
**

我们将得到一个稀疏的矩阵C，其中只有K个候选图片与目标图片比较，极大的减少了计算成本。整体过程如图2所示。



**图12.** **候选图片过滤**

将选出的K候选图片与目标图片进行比较，然后选出最佳候选图片，即确定了跟踪目标，实现跟踪。跟踪结果如图3所示。



**图13.** **基于多数据融合车辆跟踪结果**

## 5.3基于多模态数据融合的图文检索应用

在我们的日常生活中，我们经常使用搜索引擎，对于我们搜索的问题，通常是文字表达的，当我们搜索某一事物，却不知道叫什么名字，那么如何向搜索引擎表达我的问题呢？图片，用图片搜索。之所以能用图片搜索是因为，搜索引擎将文本大数据和图片大数据相融合，再将这两种数据融合分析后，文本与图片便联系了起来，可以搜索图片得出文本，也可以搜索文本得出图片。

下面以Goolge搜索引擎为例：

这里我们做了一个尝试，将 Anne Hathaway 的这幅图片作为目标图片，将其拖到Google 图片搜索引擎中去搜索。



**图14.** Anne Hathaway

然后我们得到了搜索结果：



**图15. Google 图片搜索的结果**

可以看出，检索结果都相当精确，包括给出了图片出自的网址，Google 图片搜索甚至给出了图片上人物姓名的猜测，虽然不是完全正确但也是相当了不起了。

   计算机看到的图片不仅仅是成百上千的像素，其需要从中分析颜色，亮度，从而进行图片的特征提取，以目前最具实力的 Google 为例，Google 采用了超大规模的实时并行计算。自 2010 年以来 Google 搜索引擎发生了重大变革，将其搜索迁移到了新的软件平台 Caffeine 使得搜索更加高效，其可以对来自互联网的几百亿张图片进行实时的图像特征匹配，从而更加快速的反馈用户相关的搜索结果。这些都需要大数据背景下机器学习等技术作为坚实的后盾。而数据挖掘技术又是大数据背景下机器学习应用的另一个重要领域。谁能尽早的向数据“掘金”，谁就能在今后的激烈竞争中先拔头筹。

# 6. 总结与数据融合技术的展望

## 6.1总结

本论文是一篇大数据集成与融合的综述的文章。主要是通过介绍大数据的集成与融合相关的技术与方法，典型的案例，让人们更好地去了解大数据的融合与集成。本文工作主要包括以下个部分：

1. 对于结构化数据的集成与融合，介绍了数据仓库ETL技术，以及基于Hadoop架构的大数据融合框架。
2. 对于非结构化的数据的集成与融合，主要介绍文本与图像这两种不同模态的非结构化的数据融合方法与步骤。通过此来了解大数据下非结构化数据的融合。
3. 针对当下人工智能的浪潮，介绍了传感器数据融合的方法。
4. 对于以上的三种数据融合与集成的方式，通过典型的案例(智慧城市，车辆跟踪，图文检索)进一步分析数据融合与集成如何融合到实际的生活当中。

## 6.2数据融合技术的展望

数据融合理论虽然在很多领域得到了广泛的应用，但还存在许多需解决问题，仍需要进行深入的研究。今后的研究工作应着重围绕如下几方面展开[9]。  
 (1)开发新的不确定性融合推理方法和容错能力强、实时性好的高效融合算法，其中基于人工智能的融合方法研究是融合理论研究的热点，具有巨大潜力。  
 (2)开发数据融合系统的数据库和高速并行检索推理机制。数据融合系统需要处理大量的多种形式的数据，对其进行合理的储存、维护、查询等各种操作，并常有实时性要求，所以需要建立高效、实时、使用方便、性能可靠的数据库管理系统和检索推理机制。  
 (3)数据融合的工程化方法、评估方法的研究。理论到实际的应用必须有一套合理的实现方法，同时还要对建立好的融合系统进行性能的判断。  
 (4)数据融合技术在特定领域应用的研究。不同领域对数据融合理论的应用有着不同的方法和特点，需要进行专门的研究，特别是在目标识别领域的应用得到了广泛的重视。  
 (5)多传感器系统传感器的管理。多传感器组成了多传感器系统的互补体系，必须对它们进行有效的管理，包括传感器的优先级确定、失效处理、口径选择、配置等等。以便获得最优的数据采集性能，得到最佳的融合效果。

# 参考文献：

1. 数据融合技术的发展与应用现状分析[J].2016-07-26
2. 工业大数据应用如何进行“集成应用”[N].通信世界网.2016-05-11
3. 张义，杜博文.智慧城市多模式数据融合模型.[J].北京航空航天大学学报.2016.12
4. L1 H，XUE L，ZHU Y，et a1．The application and implementation research of smart city in China[C]∥2012 InternationalConference on System Science and Engineering(ICSSE)．Pisca—taway，NJ：IEEE Press，2012：288-292．
5. COLLINS J P．Sailing on an ocean of 0 s and 1 s[J]．Science，2010．327(5972)：1455-1456．
6. 数字冰雹：城市运行监控.新浪科技[N].2016-05-25
7. NAM T，PARDO T A．Conceptualizing smart city with dimen—sions oftechnology，people，and institutions[C]∥InternationalDigital Government Research Conference：Digital GovernmentInnovation inChallenging Times．New York：ACM，201 1：282-291．
8. SCHOEPFLIN T N，DAILEY D J．Dynamic camera calibrationof roadside trafficmanagement cameras for vehicle speed esti·mation[J]．IEEE Transactions on Intelligent TransportationSystems，2003，4(2)：90—98．
9. 张辉.数据融合技术的发展[J].2016-06-12

[1]张辉.数据融合技术的发展

[2]Yu Zheng, “Methodologies for Cross-Domain Data Fusion: An Overview.”, IEEE TRANSACTIONS ON BIG DATA, VOL. 1, NO. 1, JANUARY-MARCH 2015

[3]张水平,程超,王蓉,张凤琴,陈爱网.一种应用于DaaS的物化视图候选集生成算法[J].计算机应用研究, 2012,(11)