第 5 章　机器智能的研究方法之三：行为模拟

本章学习目标

- 掌握机器感知的基本原理，了解信息传感、信息融合和物联网的基本含义。
- 熟练掌握文本分类的基本概念，了解文本分类的主要方法。
- 了解 Brooks 关于感知-动作系统研究的基本内容。
- 掌握机器学习的基本概念，了解机器学习的主要方法。
- 了解计算智能的基本概念。
- 熟练掌握行为模拟方法的特点和局限性。

本章主要介绍机器智能的第三种研究学派——行为模拟（也常被称为行为主义）的基本内容。行为模拟的主要特点就是关注智能行为表现的先后关系，而不关注其内在的结构或过程，因此其研究内容相比结构模拟的神经网络和功能模拟的专家系统而言显得更加简洁。简言之，行为模拟就是先观察情况，再做出判断，最后做出行动，即先感知，再判断，后动作。据此，本章先向读者介绍负责观察情况的机器感知，再介绍对观察到的情况做出判断的模式分类，然后介绍根据感知判断结果做出模拟智能行为动作的感知-动作系统。接着介绍与感知-动作系统密切相关的两个研究领域，一是通过学习自动扩展系统所需知识的机器学习，二是不同于传统人工智能概念的计算智能新概念，这同时可以作为是对第 1 章中有关“机器智能”命名缘由的一次呼应。最后分析行为模拟方法的主要特点与局限性。

5.1　机 器 感 知

5.1.1　机器感知基本原理

智能行为的首要特点就是要能适应环境的变化，因此，及时准确地获知环境变化的相关信息是机器智能必须具备的基础能力，而信息获取的首要环节就是信息感知，即感知客体事物运动的状态及其变化的方式。或者说是通过客体事物与感知者的相互作用，把事物的本体论信息转变为感知者的第一类认识论信息。实际上，机器感知只“感受”到了事物运动状态及状态变化方式的形式，并不“知晓”事物运动状态及其变化方式的逻辑含义和效用价值[1]102。因此，机器感知的输出结果只是语法信息，而没有语义信息或语用信息。

研究结果表明，信息感知的基本机制在于要有某种组织或器官（在人工系统场合下则是某种器件或系统）能够灵敏地感受到所关注的事物的运动状态及其变化方式。也就是说，要有某种组织或器件能够在这种事物运动状态及其变化方式的刺激下产生相应的响应。而且，这种刺激与响应之间的关系应当满足一定的条件，例如，具有一定的敏感域、敏感度、保真度。如果用 u 表示感知系统的输入刺激，用 v 表示感知系统的输出响应，用 U 表示感知系统的敏感域，用 V 表示感知系统输出响应的动态范围，那么，所希望的感知系统的输入输出关系可以表述为[1]103

$$R = \{f \mid v = f(u), u = f^{-1}(v), u \in U, v \in V\}$$

上式所规定的是一类“完全的一一对应”的互逆函数的集合。在大多数实际情况下，可以允许丢失一些非本质的信息。这时，“完全一一对应”的互逆函数关系可以有一定的放松，于是上式可以退化为

$$R = \{f \mid v = f(u), u' = g(v), d(u,u') < \varepsilon, u \in U, v \in V\}$$

这里，函数 g 不一定正好是函数 f 的逆函数，$d(u,u')$ 是 u 与 u' 之间的差异测度，ε 是这种差异测度的允许值。

分析和实验都表明，虽然人的感觉器官在感知外部事物的信息方面具有十分精巧的工作机制，但同时也存在一些天然的缺陷，主要表现在：敏感域有限；敏感度有限；分辨力不高。因此，迫切需要研究机器感知的科学和技术，制造出性能优异的机器感知系统，来辅助人类扩展自身感知信息的能力。虽然待感知的信息和相应的机器感知系统种类繁多，但它们的基本原理和实现技术却是遵循着共同的原则。机器感知信息的基本原理如下[1]105：

（1）作为事物运动状态和状态变化方式的本体论信息，可以通过与其他事物的相互作用而被后者所感知，条件是后者对前者的运动状态及其变化方式敏感。

（2）信息感知的实质是本体论信息向第一类认识论信息的语法信息的转换，它的技术本质是事物运动状态及其变化方式的载体转换。

（3）由于具体事物运动状态及其变化速度的有限性，信息感知在理论上有可能做到不丢失本体论信息的基本信息。

（4）但是，无论怎样精巧地设计信息感知系统，它的第一类认识论信息的信息量（输出信息量）都不可能超过本体论信息的信息量（输入信息量）。这是信息获取领域的信息不增原理。

也可以对机器感知信息的过程和实现技术做出如下的归纳[1]104：

（1）信息感知系统通常由敏感单元和表示单元构成。

（2）应当针对不同事物的运动状态及其变化方式研制出不同的敏感单元，以获得尽可能高的敏感度、分辨率和保真度。

（3）表示单元往往都是把敏感单元的输出响应转换为便于处理的电信号或光信号的表示形式。从这个意义上来看，表示单元可以被理解为是一种“换能器”，即完成非电（光）信号转换成为电（光）信号。

目前已经研制成功了不同种类、性能良好的传感器。例如，光敏电阻、光电管、太阳能电池、红外线传感器、紫外线传感器等都属于光信号敏感的传感器，它们共同的工作机理是能够灵敏地感知光参量的状态及其变化，并转换为电参量的状态及其变化；各种麦克风、助听

器、拾音器、医用心音计、超声波诊断传感器等都属于声音信号敏感的传感器，它们的工作机理是能够灵敏地检测到声振动参量的状态及其变化，并转换为电参量状态及其变化；气体传感器、湿度传感器、离子传感器、生物传感器则都属于对化学物质敏感的化学传感器，它们能够感知某些化学物质成分的数量状态及其变化，并转换为相应电参量的状态及其变化；电阻温度计、热敏电阻和温差电偶都是温度敏感的传感器，它们能把所处环境中的温度状态及其变化转换为电流输出大小的变化，等等。

事实上，单一传感器的工作能力远远不能满足人们对于空间、时间、可靠性等多方面的要求。随着更多应用向更加复杂的情况发展，为了全面准确地感知环境信息，必须采用方便布控、可以协同工作的多种类型的传感器，并对多种传感器的输出结果进行适当的综合或融合。由此，就形成了两个重要的科研领域：无线传感器网络与信息融合。其中，无线传感器网络更是与新兴的物联网研究密切相关。下面分别介绍它们的基本概念，希望读者可以从中体会到机器感知的研究意义与应用价值，拓宽科技视野。

5.1.2　无线传感器网络基本概念

近年来随着科学技术的飞速发展，已经出现了许多新的信息获取以及处理模式，无线传感器网络(Wireless Sensor Network，WSN)就是其中最引人注目的一种。得益于微电子技术、无线通信、计算技术、自动控制和人工智能等的共同进步，传感器向着体积小、功耗低、功能多、智能性的方向快速发展，出现了能够在微小体积内集成信息采集、数据处理和无线通信等多种功能的新型传感器。这些新型传感器价格低廉，可以根据需要随机散布在各种环境中，通过无线通信技术自组织地构成网络，其中的节点就是传感器。

具体而言，在微电子技术领域，微机电系统(Micro-Electro-Mechanism System，MEMS)的迅速发展奠定了设计和实现片上系统(System on Chip，SoC)的基础，使得将多种传感器集成为一体，制造小型化、低成本、多功能的传感器节点成为可能。而大量的 MEMS 传感器节点只有通过低功耗的无线电通信技术连成网络才能够发挥其整体的综合作用。在计算技术领域，更小、更廉价的低功耗计算设备代表的“后 PC 时代”冲破了传统台式计算机和高性能服务器的设计模式，普及的网络化带来了难以估量的计算处理能力。在通信方式上，多种无线通信技术的发展为微传感器间通信提供了多种选择，尤其是以 IEEE 802.15.4 为代表的短距离无线电通信标准的出现，无疑为无线传感器网络的发展奠定了坚实的基础。同时，具有群体智能的自治系统的行为实现和控制是自动控制和人工智能领域的前沿研究内容，从而为无线传感器网络的智能性提供了有力的技术支持[2]1。

无线传感器网络就是由部署在监测区域内大量的廉价微型传感器节点组成，通过无线通信方式形成的一个多跳的自组织的网络系统，其目的是协作地感知、采集和处理网络覆盖区域中感知对象的信息，并发送给观察者。传感器、感知对象和观察者构成了传感器网络的 3 个要素。如果说 Internet 构成了逻辑上的信息世界，改变了人与人之间的沟通方式，那么，无线传感器网络就是将逻辑上的信息世界与客观上的物理世界融合在一起，从而极大地扩展了人类认识世界的能力[3]1。

借助于传感器节点中内置的形式多样的传感器，无线传感器网络能够感知到所在周边环境中的热、红外、声呐、雷达和地震波信号，从而探测包括温度、湿度、噪声、光强度、压力、

土壤成分、移动物体的大小、速度和方向等众多物理现象，并通过无线通信传送信息。随着微电子技术、微机械加工技术、高能电池技术的发展，传感器节点的体积将越来越小，功耗越来越少，价格也会越来越低，而计算能力将越来越强，最终达到所谓“智能尘埃”的水平。这些廉价的微小节点可以大规模地部署到各类特殊环境中，对目标形成全方位的连续监测[2]1。

5.1.3 信息融合基本概念

当今的信息化社会已经布满了各种各样的信息传感器以及传感器网络，它们所获得的各类信息必须有机地融合在一起才能真正发挥出理想的作用。于是，信息融合技术就成为信息技术的前沿领域之一。

所谓信息融合就是指采集并集成各种信息源、多媒体和多格式信息，从而生成完整、准确、及时和有效的综合信息。它比直接从各信息源得到的信息更完整、更简洁、更少冗余和更有用途。而信息融合技术则是协同利用多源信息，将来自多个传感器或多源的观测信息进行分析、综合处理，以获得对同一事物或目标的更客观、更本质认识的信息综合处理技术[4]40-41。

信息融合的另一种普遍说法是数据融合，但就内涵而言，信息融合更广泛、更确切、更合理、更具有概括性，因为它不仅包括数据，而且也包括了信号和知识[5]。

信息融合的基本原理和人脑综合处理信息一样，充分利用多个传感器资源（如人的眼、耳、鼻、四肢），通过对传感器及其观测信息的合理分配和使用，把多个传感器在空间和时间上的冗余或者互补信息依据某种准则来进行组合，以获得被测对象的一致性解释或描述。其基本目标是：通过数据组合而不是出现在输入信息中的任何个别元素推导出更多的信息。可见，通过信息融合所得到的信息利用了多个传感器共同或联合操作的优势，是多传感器协同作用的结果，从而能够提高传感器的有效性[6]。

不妨继续把展望的目光放得更宽更远，设想传感器节点的智能水平进一步提高，更接近人的感觉器官，例如，可以识别人类的语言、神态、形体、动作等信息，并可以将这些不同类型的信息有机地融合在一起，从而接近或达到人类的高级感知水平，这必将带来人机交互模式的变革，不再是人类适应计算机，而是机器真正与人交流、为人服务了。可以预见，这必将引起人类工作生活的方方面面都发生根本性的变化，社会也将随之极大地进步[4]40。

5.1.4 物联网基本概念

物联网是近年来出现的研究热点之一，由于其与传感器网络和智能密不可分，在此对其进行简要介绍。

物联网（Internet of Things），通俗来说就是“物物相连的网络”，是利用传感器、射频识别（Radio Frequency IDentification，RFID）、二维码等作为感知元件，需要通过基础网络来实现物与物、人与物的互联[7]。1999 年，麻省理工学院 Kevin Ashton 教授首次提出的物联网概念以标示为特征，将 RFID 与传感器技术结合，应用于物品中构成物联网雏形[8]。2005 年 11 月 17 日，在威尼斯举行的信息社会世界峰会（World Summit of Information Society，WSIS）上，国际电信联盟（ITU）发布的《ITU 互联网报告 2005：物联网》指出，物联网是通过

RFID 和智能计算等技术实现全世界设备互联的网络。形成了以互联为特征的物联网概念，无所不在的“物联网”通信时代即将来临，世界上所有的物体，从轮胎到牙刷、从房屋到纸巾，都可以通过因特网主动进行信息/数据交换。2009 年 2 月 24 日，IBM 公司大中华区首席执行官钱大群在 2009 IBM 论坛上公布了名为“智慧的地球”的最新策略。物联网概念转而以智能服务为特征，IBM 公司提出：把传感器设备安装到电网、铁路、桥梁、隧道、供水系统、大坝、油气管道等各种物体中，并且普遍链接形成网络，即“物联网”[7]。

2009 年 8 月上旬，温家宝总理在无锡视察时指出，“要在激烈的国际竞争中迅速建立中国的传感信息中心或‘感知中国’中心”。此后，物联网在中国受到了极大的关注，其程度是在美国、欧盟以及其他各国不可比拟的。物联网这个词在中文习惯里比“感知中国”更朗朗上口，而且与互联网很对应，所以成了更被大众接受的说法。物联网的概念与其说是一个外来概念，不如说它已经是一个“中国创造”的概念，它的覆盖范围与时俱进，已经超越了 1999 年 Ashton 教授和 2005 年 ITU 报告所指的范围，物联网已被贴上“中国式”标签。

“中国式”物联网是指将无处不在的末端设备和设施，包括具备“内在智能”的传感器、移动终端、工业系统、楼控系统、家庭智能设施、视频监控系统等，和“外在使能”的，如贴上 RFID 的各种资产、携带无线终端的个人与车辆等等“智能化物件或动物”或“智能尘埃”，通过各种无线/有线的长距离/短距离通信网络实现互联互通、应用大集成以及基于云计算的 SaaS 营运等模式，在内网、专网/互联网环境下，采用适当的信息安全保障机制，提供安全可控乃至个性化的实时在线监测、定位追溯、报警联动、调度指挥、预案管理、远程控制、安全防范、远程维保、在线升级、统计报表、决策支持、领导桌面等管理和服务功能，实现对万物的高效、节能、安全、环保的“管、控、营”一体化[9]。

可见，“智慧的地球”和“中国式”物联网都在强调智能服务和智能决策，这又一次很好地体现出了智能科学技术的必要性和重要性。

实际上，物联网被认为是继 WWW 和移动互联网之后的互联网革命的第三波。物联网相较于传统的互联网和传感网具有以下几个优点：更透彻的感知，即利用任何可以随时随地感知、测量、捕获和传递信息的设备、系统或流程，便于立即采取应对措施和进行长期规划；更全面的互联互通，即将个人电子设备、组织和政府信息系统中储存的信息交互和共享，从而对环境和业务状况进行实时监控；更深入的智能化，即使用数据挖掘和分析工具、科学模型和功能强大的运算系统处理复杂的数据分析、汇总和计算，整合和分析海量的跨地域、跨行业的信息，以更好地支持决策和行动[7]。

5.2　模 式 分 类

在通过机器感知获得了大量信息之后，接着就需要对这些信息进行判断或识别，这个任务就是模式分类，也可称之为模式识别。基于第 2 章中关于信息概念的介绍，机器感知到的信息可以具体分为语法信息、语义信息和语用信息，其中以语法信息最为简单，因此，最简单的模式分类问题就是针对语法信息的处理，它重点关注的是信息的形式关系，基本工作原理就是基于形式信息的比较。

为了使这种基于形式的比较能够有效地实现，往往不是原封不动地把待识别的信息与

相应的模板进行比较，而是把能够表征这个语法信息的一组形式化参量（称为特征）提取出来形成模式，再与相应的类别特征模板进行比较，根据它们之间的匹配情况来判断信息所归属的模式类别[1]107-108。

5.2.1 模式分类基本概念

实际上，类别是大自然和社会生活中的一种固有现象，正如一句俗语所说，“物以类聚，人以群分”。自然和社会中处处都可见到相似的物体聚集在一起，而不相似的物体往往相互分开。例如，一种环境中往往会聚集很多同类的动物或植物，图书馆中的书籍都是同类的摆放在一起以方便查找，人群聚会中常常是兴趣相投的人们组成小团体，不同兴趣的团体则相互分离，等等。人类经过长期积累形成的知识体系中包含了大量的概念，它们正是人类对自然和社会中各种类别存在的最直接的认知结果。

简单说来，模式分类就是根据分类标准判断输入模式的所属类别，可以看成是按照某种既定的分类体系给待分类的模式对象贴上不同类别的标签。分类现象无处不在，因此模式分类的应用也是无处不在，感兴趣的读者也可以尝试用模式分类的眼光来重新审视所在的世界。

当然，自然和社会中的模式是千变万化的，模式分类的方法相应地也有很多。一般来说，规则的模式分类问题可以用数学方法进行严格的描述和分析，如在白噪声背景中数字信号的识别（检测）问题。白噪声具有明确的统计特性，数字信号（信息）具有规则的形式（0或1），且状态转换的方式也服从某种统计规则，因此，它的识别过程可以用概率论方法进行定量的分析。这就是模式识别理论中的统计方法或统计决策理论方法。但是，有许多模式识别问题却不可能完全用解析的方法求解，而必须借助于启发式的算法进行推断，如模式识别理论中的句法方法或语言学方法。此外，通过大量示例训练的方法，在模式的形式或模式的特征与模式的类属之间建立某种非线性的映射关系，然后利用这种映射关系对未知模式进行分类，也是一种有用的模式识别方法，这就是基于神经网络的方法[1]108。

模式分类的应用极其广泛，典型的应用有文本分类、文字识别、语音识别、图像识别、错误诊断等等。还有很多应用尽管表面看起来各不相同，但其核心技术就是模式分类，例如信息检索、信息抽取、话题检测与跟踪、趋势预测、信息推荐、垃圾信息过滤等等。下面仅以文本分类为例来展开说明。

5.2.2 文本分类基本概念

随着互联网技术的迅速发展和普及，如何对浩如烟海的文献、资料和数据（很大一部分是文本）进行自动分类、组织和管理，已经成为一个具有重要用途的研究课题。文本自动分类简称文本分类（text categorization），是模式识别与自然语言处理密切结合的研究课题。传统的文本分类是基于文本内容的，研究如何将文本自动划分成政治的、经济的、军事的、体育的、娱乐的等各种类型[10]416。

文本分类就是指根据预先定义的主题类别，按照一定的规则将文档集合中未知类别的文本自动确定一个类别[11,12]。通常，文本分类系统的输入是需要进行分类处理的大量文本，而系统的输出是与文本关联的类别。简单地说，文本分类就是对文档标以合适的类标签。

从数学的角度来看,文本分类是一个映射过程,它将未标明类别的文本映射到现有类别中,该映射可以是一一映射,也可以是一对多映射,因为一篇文本可以与多个类别相关联[13]。可见,一个文本分类系统不仅是一个自然语言处理系统,也是一个典型的模式识别系统[10]416。

文本分类问题自身也可以有不同的分类方法。最简单的是可以根据类别数目的不同划分成两种:一种是以单类别为基础的两分类问题,其类别体系仅仅包括两个互补类,即一篇文本属于或不属于某个主题类别;另一种是涉及多个类别的多分类问题,其类别体系通常由 3 个或者 3 个以上的类别构成,一篇文本可以属于其中某一个或者多个类别。其中,两分类问题是多分类问题的基础,从实现方法的角度看,多分类问题常常可以通过拆分方便地转化成多个两分类问题来实现,当然也有很多直接处理多个类别的实现方法。

此外,文本分类问题也有其他的划分方法,例如,可以根据分类结果中是否存在兼类现象划分成两种:一种是一个文本只属于一个类别,可以称为单标签问题;另一种则允许出现兼类现象,一个文本可以同时属于多个类别,称之为多标签问题。

显而易见,预先定义的分类体系是文本分类问题不可或缺的基础,而分类体系的制定则完全取决于目标应用的需求。例如,门户新闻网站需要的分类体系通常是按照新闻内容的主题划分的,如政治、经济、军事等,学术论文资源库的分类体系则往往是按照学科划分的,如哲学、数学、物理、医学等。事实上,分类体系一般由人工通过需求分析制定,大多数采用层次结构,就是一棵分类树。已有的一些比较著名的分类体系包括中图分类法、Reuters 语料分类体系、Yahoo!分类目录等。

文本分类的应用也极其广泛,例如垃圾邮件判定(两分类问题,垃圾邮件/非垃圾邮件)、个性化信息推荐(两分类问题,推荐/不推荐)、信息检索(两分类问题,相关/不相关)、按照栏目分类的新闻出版(多分类问题,如政治、体育、军事等)、产品评论的情感分析(多分类问题,如喜欢、讨厌、中立等)、事件检测与跟踪(多分类问题,事件 1、事件 2 等)。

经过 50 多年的研究,目前比较成熟的文本分类系统已经有很多种了,大致可以分为两大类,即基于知识工程的分类系统和基于机器学习的分类系统。在 20 世纪 80 年代,文本分类系统以知识工程的方法为主,根据领域专家对给定文本集合的分类经验,人工提取出一组逻辑规则,作为计算机文本分类的依据,然后分析这些系统的技术特点和性能。进入 20 世纪 90 年代以后,基于统计机器学习的文本分类方法日益受到重视,这种方法在准确率和稳定性方面具有明显的优势。系统使用训练样本进行特征选择和分类器参数训练,根据选择的特征对待分类的输入样本进行形式化,然后输入到分类器进行类别判定,最终得到输入样本的类别[10]417。

可见,基于知识工程的分类系统就是由人工总结规则并用计算机程序来实现。例如,若文本中存在词语“篮球”和“联赛”,那么可以归为体育类文本。其优点是结果很容易被人理解,但它的缺点也是显而易见的:人工制定规则费时费力,而且又难以保证所有规则的一致性和准确性,因为制定分类规则的专家可能有多位,每人的理解或许存在偏差,甚至有时候专家也是依靠想象的,缺乏足够真实语料的经验。正是为了克服这些缺点,基于机器学习的分类系统是从训练语料中自动学习分类规则。与基于知识工程的分类系统相比,其优点是速度快,准确率相对较高,且来源于真实文本,可信度较高。但它也有缺点:分类结果可能

不够直观，难以被人理解，因为其所学习到的规则往往是一些复杂的数学表达式。

简言之，基于知识工程的分类系统通过得到某些规则来指导分类，而基于机器学习的分类系统则是通过计算得到一些数学表达式来指导分类。归根结底，它们在本质上是相通的，都是想得到某种规律性的知识来指导分类，通过统计机器学习得到的数学表达式也可以看作某种隐式的规则。

在目前的文本分类系统中，统计机器学习方法还是占据了主流地位。使用统计方法进行文本分类通常包括两大步骤：步骤 1 是训练，就是从训练样本中学习分类的规律性知识；步骤 2 是分类或测试，就是根据步骤 1 学习到的知识对新输入的文本进行类别判定。同时，无论是训练还是分类测试，都不是直接使用原始文本，而是要分析出文本的某些特征，即特征抽取，然后把文本变成这些特征的某种适宜处理的表示形式，即特征表示。使用统计方法进行文本分类的过程可用图 5.1 表示。

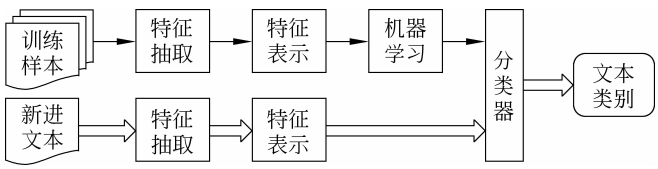

图 5.1　统计文本分类过程

图 5.1 中，单线箭头表示的就是步骤 1 训练，训练结果就反映在分类器中。空心箭头表示的是步骤 2 分类或测试，其对新进文本的特征抽取和特征表示方法与步骤 1 是一致的，使用的就是步骤 1 得出的分类器，最终输出的是文本类别判定结果。

5.2.3　文本特征抽取与表示

一个文本表现为一个由文字和标点符号组成的字符串，由字或字符组成词，由词组成短语，进而形成句、段、节、章、篇的结构。要使计算机能够高效地处理真实文本，就必须找到一种理想的形式化表示方法，这种表示一方面要能够真实地反映文档的内容（主题、领域或结构等），另一方面要有对不同文档的区分能力。目前文本表示通常采用向量空间模型（Vector Space Model，VSM）[10]417-418，[14]。

为了方便进行文本特征抽取，通常首先要对原始文本进行预处理，以便得到文本的基本内容单元——词的特征信息。因而它的主要任务是进行分词和去停用词。分词就是将文本字符串切分开来，得到一个个的词，其具体实现方法往往与文本所使用的语言有关。例如，对英语而言，分词可以直接依据单词之间天然存在的空格进行；而对汉语而言，文本字符串中则没有空格，因此需要专门的中文分词系统，也正是这个原因，中文分词算法的研究作为中文信息处理的基础一直是我国学者的研究重点之一。分词之后所得到的词汇并不都是必要的。停用词就是指在各类文本中都频繁出现，因而被认为带有很少的有助于分类任何信息的代词、介词、连词等高频词[15]。为了减少停用词所带来的负面影响，人们通常要进行去停用词的处理，常用的方法是构造一个停用词表，并据此将文本中的停用词去除。

近年来，网络文本日益成为文本分类的研究对象，而其中除了文本内容之外，还包含很多其他信息，如网络页面中的各种标签语言、脚本代码等。这些在文本分类时也都是不必要

的干扰,因此网页文本预处理中往往还需要去除这些噪声以获得干净的纯文本。但是这些噪声变化无穷,难以彻底清除,于是又产生了一个专门研究网页去噪的课题任务。

此外,为了获得更多的词语信息甚至概念语义信息,研究者们也提出了更深入的预处理步骤,例如针对英语的词根还原(stemming)、词性标注、短语识别、同义概念识别、相关概念聚类等等。相关研究表明,用概念代替单个词可以在一定程度上解决自然语言的歧义性和多样性给特征向量带来的噪声问题,有利于提高文本分类的效果[10]419。

下面给出向量空间模型相关的两个基本概念[10]418。

(1) 项/特征项(term / feature term): 是指最小的不可分的语言单元,可以是字、词、词根、词组或短语等。一个文档的内容被看成是它含有的特征项所组成的集合,表示为

$$\text{Document} = D(t_1, t_2, \cdots, t_n), \quad \text{其中}, t_k \text{ 是特征项}, 1 \leqslant k \leqslant n。$$

(2) 项的权重(term weight): 对于含有 n 个特征项的文档 $D(t_1, t_2, \cdots, t_n)$,每一特征项 t_k 都依据一定的原则被赋予一个权重 w_k,表示它们在文档中的重要程度。这样一个文档 D 可用它含有的特征项及其所对应的权重来表示:

$$D = D(t_1, w_1; t_2, w_2; \cdots; t_n, w_n)$$

简记为 $D=D(w_1, w_2, \cdots, w_n)$,其中 w_k就是特征项 t_k的权重,$1 \leqslant k \leqslant n$。这个权重也可以看作文档在各个特征维度上的投影。

于是,一篇文档可以看成是 n 维特征空间中的一个向量,这就是向量空间模型的由来。需要注意的是,该模型认为各个特征项之间是相互独立并且没有先后顺序关系的。尽管这与自然语言文本事实并不完全相符,但是这种简化直接促成了文本计算模型的可实现性大大增强。此后也有很多研究者对向量空间模型提出了多种改进方法,感兴趣的读者可自行查阅相关文献。

具体到特征项的选择,可以根据可用的文本预处理工具以及文本分类任务的需求而定。下面给出几个不同特征项的例子,以便帮助读者更好地理解掌握。

- 字特征项: 北、京、邮、电、大、学
- 词特征项: 北京、邮电、大学
- 短语特征项: 北邮、北邮图书馆
- 概念: 北京邮电大学
- 同义词: 北京邮电大学/北邮/BUPT
- 相关词: 北京邮电大学/邮电高校

在构造文档向量的基础上,文档之间的内容相关程度就可以方便地用向量之间的相似性度量来实现了,如图 5.2 所示。

图 5.2 中,虚线箭头表示各个特征项维度,所有特征项共同形成了 n 维特征空间。一篇文档就是其中的一个向量,图中的两个实线箭头表示任意两个不同的文档向量,分别用 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 表示,即:

$$\boldsymbol{D}_1 = \boldsymbol{D}_1(w_{11}, w_{12}, \cdots, w_{1n}), \quad \boldsymbol{D}_2 = \boldsymbol{D}_2(w_{21}, w_{22}, \cdots, w_{2n})$$

令 $\text{Sim}(\boldsymbol{D}_1, \boldsymbol{D}_2)$ 表示这两个文档之间的相似性,θ 表示这两个文档向量之间的夹角,那么,可以借助于 n 维空间中两个向量

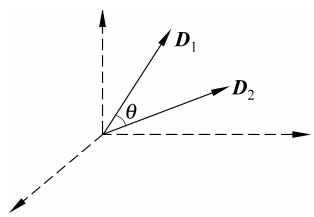

图 5.2　向量空间模型示意图

之间的某种距离来表示文档间的相似性。数学上计算向量距离的方法有很多，常用的有内积计算、夹角余弦 Cosine 计算、Dice 计算、Jaccard 计算等，使用这些方法计算 $\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)$ 的公式列举如下。

内积计算：$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\boldsymbol{D}_1\cdot\boldsymbol{D}_2=\sum_{k=1}^{n}w_{1k}\times w_{2k}$

Cosine 计算：$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\cos\theta=\dfrac{\boldsymbol{D}_1\cdot\boldsymbol{D}_2}{\|\boldsymbol{D}_1\|\times\|\boldsymbol{D}_2\|}=\dfrac{\sum_{k=1}^{n}w_{1k}\times w_{2k}}{\sqrt{\sum_{k=1}^{n}w_{1k}^2\sum_{k=1}^{n}w_{2k}^2}}$

Dice 计算：$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\dfrac{2\times\boldsymbol{D}_1\times\boldsymbol{D}_2}{\|\boldsymbol{D}_1\|^2+\|\boldsymbol{D}_2\|^2}=\dfrac{2\sum_{k=1}^{n}w_{1k}\times w_{2k}}{\sum_{k=1}^{n}w_{1k}^2+\sum_{k=1}^{n}w_{2k}^2}$

Jaccard 计算：$\mathrm{Sim}(\boldsymbol{D}_1,\boldsymbol{D}_2)=\dfrac{\boldsymbol{D}_1\cdot\boldsymbol{D}_2}{\|\boldsymbol{D}_1\|^2+\|\boldsymbol{D}_2\|^2-\boldsymbol{D}_1\cdot\boldsymbol{D}_2}$

$$=\frac{\sum_{k=1}^{n}w_{1k}\times w_{2k}}{\sum_{k=1}^{n}w_{1k}^2+\sum_{k=1}^{n}w_{2k}^2-\sum_{k=1}^{n}w_{1k}\times w_{2k}}$$

可见，文档向量中特征项的权重是计算文档相似性的关键。最基础且最容易理解的项权重计算方法为布尔权重(boolean weighting)，即

$$w_{ij}=\begin{cases}1, & \mathrm{TF}_{ij}>0\\ 0, & \mathrm{TF}_{ij}=0\end{cases}$$

其中，TF 代表的是项频率，取自英文 Term Frequency，TF_{ij} 表示的是特征项 t_i 在文档 D_j 中出现的次数。可见，布尔权重的主旨就是：如果文本中出现了某特征项，那么文本向量中该特征项的权重就为 1，否则为 0。

然而，也正是由于布尔权重计算方法比较简单，它的缺点也是显而易见的，即无法区分不同特征项对文本作用的不同。于是研究者们继而提出了很多其他的权重计算方法，如倒排文档频率 IDF、TF-IDF 型权重、TFC 型权重、熵权重等等[10]422-423。其中最经典也是使用最为普遍的就是 TF-IDF 型权重，即

$$w_{ij}=\mathrm{TF}_{ij}\times\mathrm{IDF}_i=\mathrm{TF}_{ij}\times\log\frac{N}{\mathrm{DF}_i}$$

其中，DF 代表的是文档频率，取自英文 Document Frequency，IDF 代表的是倒排文档频率，取自英文 Inverse Document Frequency。N 表示所有文档的数目，DF_i 表示的是所有文档中出现特征项 t_i 的文档数目，IDF_i 就是特征项 t_i 的倒排文档频率。

可见，TF-IDF 型权重与特征项在文档中的出现频率成正比，且与在所有文档中出现该特征项的文档数目成反比。可以将其与文档中特征项的使用情况关联起来理解。一般地，一篇文档中的重要特征项会被反复提及，于是其频率也就更高。而对于大量文档集合而言，如果某一特征项在其中大多数文档甚至所有文档中都出现过，那么该特征项对于区分这些文档是没有帮助的；相反，如果某一特征项仅在一小部分文档中出现过，那么该特征对于区