第3章

信息网络技术与数据获取

大数据分析的应用离不开数据本身，目前大数据分析的数据主要来源于互联网的第三方数据。互联网的基础是信息网络技术，本章从信息网络技术出发，介绍信息网络技术的基础知识、大数据时代下的信息网络特点，明确信息网络的组成与资源共享方式，接下来将信息网络技术与大数据分析应用结合，介绍如何通过 Python 程序设计语言构建网络爬虫实现大数据分析过程中的数据获取，从而充分利用互联网中的海量信息和数据进行分析，获取更有价值的关键决策信息。

3.1 信息网络概述

3.1.1 网络的结构

网络从字面上理解，泛指网状的东西或网状系统，由若干节点和连接这些节点的链路构成。从动物界中蜘蛛张网捕虫到封建等级制度，再到社交网络“六度空间理论”，网络无处不在，而大数据时代，“万物皆互联，无处不计算”的特性使得网络的边界和范围更加深化。

在网络的构成要素中，节点既可以是现实世界的人或物，也可以是虚拟世界的逻辑节点。而反映节点之间关联关系的边则决定了网络的不同形态，社会网络、交通网络、信息网络、组织网络等都是其不同展示形式的直观体现。例如，人与人之间的社交联系就构成了一张属于自己个人的社交关系图，个人在自己的社交网络中进行沟通交流、社会协作、社交维护，同时借助网络的扩展特性不断延伸其社交范围。再者社会结构网络受特定生产力水平下的代表统治阶层利益的管理制度和管理跨度所限制，西欧封建等级制度则是这一结构最直观的体现，封建主以土地关系为纽带，通过层层分封、依次互为主从建立起用来协调和维护封建统治阶层利益的从上到下的金字塔式的治理网络。

美国哈佛大学心理学教授斯坦利·米尔格拉姆在 1967 年的连锁信件实验中提出“六度空间理论”，认为世界上任意两个人之间建立联系，最多只需要 6 个人。虽然在很长一段时间内，米尔格兰姆的实验结论备受争议，但其揭示一个很重要的现象：任何两个素不相识的

人,通过一定的方式总能取得联系。同时期的社会学家马克·格拉诺维特在进行“如何找工作”的调查过程中发现“弱关系”的存在及其显著作用,并在1973年的论文中,将社会关系进一步划分为“强关系”和“弱关系”,格拉诺维特认为社会网络中普遍存在的“弱关系”在我们与外界的信息交流过程中发挥重要作用,“弱关系”虽然不如“强关系”具有高度的互动性,但却有着更快速、低成本、大范围的信息传播优势。进入互联网时代,信息技术的快速发展、SNS社区(Facebook、人人网、QQ等)的出现使得“六度空间理论”和“强弱关系”在数字时代得到充分验证,相同爱好、相同兴趣、相同圈子的群体组成的弱关系网络在人们的生活中扮演着重要的作用。新科技背景下,大数据、人工智能、区块链、物联网与信息网络的融合扩展了网络的功能边界,为信息网络的发展和演化提供新的发展动力。

3.1.2 信息网络的起源

“上古结绳而治,后世圣人易之以书契”。商周时代以前,绳纹成为中国最早的文字符号,绳上的每个结代表一件事,大事结大结,小事结小结[①]。以绳和绳结构成的记事网络实现了文化信息的传承,而在商周时代盛行的钟鼎文化则是另一种认知网络的交流方式,结绳记事如图3-1所示。以绳结和钟鼎为代表的文字符号因其客观局限性远远满足不了人类对信息记录和传递的需求。商周时代以后,竹简作为早期中国文化的信息载体对华夏文明的传承和发扬起到了至关重要的作用,先秦诸子百家争鸣的文化盛况得以保存,四书五经等重要的文化典籍得以流传。而竹简作为先秦西汉时代重要的信息载体,其缺点也很明显,其制作成本昂贵、搬运不便,更多时候为少数大儒和权贵所占有,平民阶层往往难以获取。东汉永元十七年(公元105年),蔡伦在改进前人丝织品的经验基础之上,革新制造工艺,制造出成本更低、质量更优的“蔡侯纸”。造纸术的出现为知识信息提供低廉、便捷的传播渠道,打破了少数阶层对知识信息的垄断,以纸为媒介的信息网络逐步向社会各个阶层传播。

图3-1 结绳记事

工业革命时期,美国一位贫困的意大利裔移民安东尼奥·梅乌奇在移居古巴期间,研究用电击法治病时,发现声音可以以电脉冲的形式穿过钢丝,为探索其中的奥秘,梅乌奇在移居美国后开始该项研究,并于1860年向公众展示了世界上首部电话的雏形。“沃森先生,请立即过来,我需要帮助!”比梅乌奇小两岁的美国发明家亚历山大·贝尔成功通过电话传出的第一句话掀开了人类通信史上的新篇章。1888年,在德国科学家赫兹发现的电磁波基础之上,意大利人马可尼制出了无线电通信设备,并于隔年在英法两国之间发报成功。近代电信事业的发展,为快速传递信息提供了方便。从此世界各地的经济、政治和文化联系进一步加强。电话与无线通信技术成为缩短空间和时间的机器,让分隔两地的人即时联系成为现实,提高了人与人之间的信息交流的速度,促进了跨地区商业和贸易的流通。

而在人类社会进入21世纪以来,随着互联网通信技术的广泛应用,人与人之间的时空

① 事大,大结其绳;事小,小结其绳,之多少,随物众寡。——《易九家言》

距离骤然缩短，整个世界紧缩成一个“村落”，通过这个网与其他人联系起来，并且为人们的社会生活提供了方便、快捷、高效、共享的平台。如今的我们可以在家里、户外等任何场所使用手机、平板、计算机等互联网终端设备与来自世界各地的人们进行交流协作。这一改变彻底打破了人们对固定工作和学习的老旧思想，使整个世界成为互通有无的共同体。

而融合信息传感器、射频识别技术、全球定位系统、红外感应器、激光扫描器等各种装置与技术的物联网(Internet of Things，IoT)时代到来之后，万物互联正在朝人们的生活迎面走来。互联网让世界各地的人们连接起来，而物联网可以实现物与物、物与人的泛在连接，实现对物品和过程的智能化感知、识别和管理。从智慧校园的一卡通到高速公路上的 ETC 不停车收费系统，再到近几年流行的智能手环、智能手表等可穿戴设备，都是物联网运用的例子。

而当物联网遇上人工智能，AI＋IoT 已成为物联网发展的必然趋势，智能家居、自动驾驶、智慧医疗、智慧办公等创新性的场景应用正逐步改变着物理世界与数字世界的连接鸿沟。家居的“智能化”就是一个实际例子，扫地机器人、智能冰箱、洗碗机、音响等将不再是单独个体，而是成为“智慧家庭”的一部分。这些智能家庭又组成智慧社区，无数智慧社区又构成智慧城市的雏形。AI＋IoT 将给人类社会带来又一次全新变革。“万物互联”的时代，已逐渐从科幻电影中进入人们的生活，不仅人与人之间形成了网络，万事万物都在这张巨大的网络上相互联系，共同创造更加美好的生活。

3.1.3　信息网络的定义

对于信息网络，更精准的定义是将分布在不同位置的、具有独立功能的计算机，通过通信设备和通信线路连接起来，完成信息交换，以实现资源共享和协同工作的计算机集合。信息网络也可以定义为信息在社会群体中的交互，即信息网络是指由多层的信息发出点、信息传递和信息接收点组成的信息交流系统，这个系统是由个体和群体的人构成的无形的网，它能贯穿上下，联系左右，沟通内外，纵横交错，通达灵便。信息网络是信息资源开发利用和信息技术应用的基础，是信息传输、交换和共享的必要手段。只有建设先进的信息网络，才能充分发挥信息化的整体效益。信息网络也是建设“数字中国”的重要基础设施。

信息网络的呈现形式多种多样，从覆盖范围来讲可以划分如下：一是纵向网络，即从上到下贯通一气的线条型网络；二是横向网络，即不同地区、不同部门、不同单位之间的联系网络；三是延伸网络，即不受管辖范围和系统范围的限制，在有必要时，超越管理层次界限，直接与某单位甚至某个人建立信息联系；四是扩散网络，即超出本系统范围的信息网络，大多是各专业部门与社会公共事业部门建立的信息网络；五是内部网络，即组织内部的各个部门之间形成的网络。

信息网络是计算机技术、网络技术、通信技术、社会科学等多种学科紧密结合的产物。它不仅使计算机的作用范围超越了地理位置的限制，而且也大大加强了计算机本身的能力。伴随着社会群体规模的扩大和新兴技术的深入融合，物理世界的边界逐渐模糊，连接方式日趋多样性。网络所蕴含的内在机制也在发生新的变化，现代网络不仅仅是信息传播的媒介，数据时代的发展为其赋予新的特征。

1. 多样性

随着人类社会的不断进步发展，网络环境也日益复杂和多样，不同领域对于网络呈现不

同程度的需求,因此多样性成为了现代网络的突出特征。而以物理网络形成的基础设施,为社会、经济、文化领域的创新提供新的发展模式。“小群体网络”“经济共同体”“粉丝网络”等的出现更加完善和丰富了网络的内涵。

2. 共享性

现代网络的共享性是以计算机设备为载体,可以通过文字、图片、视频、链接等多种方式进行交流和资源共享。同时还实现了人与人之间点对点的传播,每一个人在网络上既是生产者也是使用者。网络的共享性不仅限于信息方面,还有硬件、软件的共享,这些共享实实在在地提升了对各种资源的利用率,有效提升了社会生产力。

3. 连接性

现代网络的技术特性打破了连接主体的时空局限性,实现了跨地区、跨空间、跨群体的沟通和交流,网络的连接性推动了市场的全球化、网络化以及无国界化。如果说互联网带来的是“人与人”“人与信息”的连接,那么在互联网基础上延伸发展的物联网则更进一步,它实现了“人与物”“物与物”之间的连接。物联网不再以“人”为单一的连接中心,人与物、物与物之间无须人的操控也可实现自主连接,它所涉及的领域包括可穿戴设备、智能家居、自动驾驶汽车、互联工厂以及智慧城市等。可以预见,以物联网为代表的网络浪潮将从根本上改变我们习以为常的生活方式,也将重构全球经济社会新格局。

4. 高速性

大数据时代,人们不单单是满足于当前网络的传输速度,而是想要追求更快、更为流畅的网络传输速度。网络技术的每一次升级换代都是以传输速度的飞跃性发展为变革焦点,从最开始的1G无线通信技术到4G、5G,传输速度经历了从KB到MB再到GB的指数级增长,网络传输的低延迟、高容量和超大规模连接的能力改变人类的生活、工作方式。根据国际电信联盟(ITU)发布的5G标准规定:单个5G基站至少提供10Gb/s的上行链路,每平方千米至少承载100万台设备,单个延迟不超过4ms,能够支持高达500km/h的设备连接而不中断。以高速性为特点的现代网络将对医疗、教育、电力、文化、工业、交通等领域带来颠覆式变革。

5. 智能化

现代网络的智能化之路经历了从全人工方式,经半自动化到全自动化再到智能化的漫长演进历程。智能化意味着网络能够处理以自然语言表述的业务意图,自动将其转化为网络策略和行为,确保网络持续、可靠地满足业务需求,且不影响其他业务的运行。应用的智能化能够有效提升业务处理效率,为人们提供更加优质的服务。例如,在一些家用电器,如微波炉、烤箱、洗衣机、吸尘器当中植入人工智能技术,能够使人们的操作更加便捷化。无人驾驶汽车将传感器物联网、移动互联网、大数据分析等技术融为一体,从而能动地满足人的出行需求。

6. 技术综合性

现代网络技术作为一个整体,具有技术综合性的特征。它在发展的过程中集中了多种优势技术,实现网络科学的发展依赖计算机、网络、通信、社会科学、数学等多种技术的实现,从而形成了现代网络所具有的强大的通信和传输能力。多种技术综合使现代网络能够同时

拥有多种方式和方法来满足不同用户的需求，使其更加多元化，有利于满足信息化时代人们的多元化和个性化需求。无论用户在何时何地都可以通过网络来实现信息、数据以及资源的传输和共享，多种技术的结合也为现代网络技术的发展提供了强大的支持力量，是现代信息网络发展过程中不可或缺的一部分。

3.2　信息网络技术

从第一部电话的问世到阿帕网的诞生，以通信网络为传播媒介的数字传播渠道逐步替代以纸为载体的物理传播渠道。通信网络高速、即时的传播特性颠覆了传统的信息传播网络，为信息的全球化交换提供了平台。原有的简单网络向复杂网络的演化，信息交换标准从杂乱无序到开放互联互通，传播速度由低速单一传输向高速智能多样化的迭代。

3.2.1　信息网络技术概述

信息网络技术是计算机技术、通信技术、网络技术相结合的产物，它将网络上分散的资源进行整合，实现资源的全面共享和有机协作，使人们能够透明地使用这些资源并按需获取所需要的信息。从一定程度上来说，信息网络技术及其应用已经成为现代化国家发展中综合国力评定的重要因素，对人们的生活、企业发展和国家政治、军事、文化等的进步起到极大的推动作用。

1957 年 10 月 4 日，苏联在拜科努尔航天中心发射了人类历史上第一颗人造地球卫星 Sputnik，鉴于美苏之间冷战爆发的阴霾，时任美国总统的艾森豪威尔正式向国会提出建立国防高级研究计划署（Defense Advanced Research Projects Agency，DARPA，也常被称为 ARPA），希望通过该机构建立一个分散的指挥系统，确保在集中的军事指挥中心受到核攻击后，全面的军事指挥系统仍能正常工作，而这些分散的指挥中心通过某种形式的信息通信网络连接起来。日后被称为“阿帕网之父”的拉里·罗伯茨提交的《资源共享的计算机网络》研究报告中提出阿帕网的构想，通过“阿帕”实现分布在不同物理区域的计算机互相连接，从而使各节点的信息共享和数据交换。并于 1969 年 11 月建立了全球第一个包交换网络——阿帕网（Advanced Research Projects Agency Network，ARPANET），两周后，包含 4 个节点的阿帕网雏形建成。在阿帕网建成之初，大部分计算机的信息交换接口相互不兼容，终端软硬件的差异迫切需要一个统一的网络传输规则体系，各节点遵循统一的网络通信协议，实现全网范围内的数据通信。1974 年，由文顿·瑟夫及同事正式发布的网络控制协议（NCP）报告中提出了“传输控制协议（Transmission Control Protocol，TCP）”和“网际协议（Internet Protocol，IP）”，即当前互联网发展的重要基石——TCP/IP。1983 年，TCP/IP 正式成为 Internet 的标准协议，这一年，也被称为 Internet 的元年。

1987 年 9 月 20 日，北京计算机应用技术研究所向德国发出的第一封电子邮件“Across the Great Wall we can reach every corner in the world.”揭开了中国人使用 Internet 的序幕。1994 年 4 月 20 日以“中科院—北京大学—清华大学”为核心的“中国国家计算机网络设施”通过美国 Sprint 公司的一条 64K 国际专线实现了与全球 Internet 的互联，标志着中国正式通向国际互联网。现如今，Internet 逐渐演变成多级结构、覆盖全球的大规模网络。

根据中国互联网络信息中心(CNNIC)发布的第47次《中国互联网络发展状况统计报告》显示,截至2020年12月,我国网民规模为9.89亿,互联网普及率达70.4%,IPv6地址数量较2019年底增长13.3%,庞大的网民构成了中国蓬勃发展的消费市场,也为数字经济发展打下了坚实的用户基础。截至2021年12月底,我国网民规模为10.32亿,较2020年增长4296万,互联网普及率达73.0%。IPv6地址数量同比增长9.4%;移动通信网络IPv6流量占比已经达到35.15%。

互联网的快速发展及延伸,加速了信息的流通与汇聚,促使各种信息资源数量呈指数增长,人们融入各种信息网络中,并从中获取各种便利。互联网在人们的生产生活中的重大意义、深远影响不言而喻,已经成为与人们的衣食住行一样不可或缺的必需品。在互联网和信息化时代的背景下,信息网络技术不断地衍生和发展,形成庞大的信息网络技术体系,它能够把网络中分散的资源融为有机整体,实现资源的全面共享和有机协作,使人们能够透明地使用资源的整体能力并按需获取信息,其中资源包括高能性计算机、存储资源、数据资源、信息资源、知识资源、大型数据库、网络、传感器等,使人们的生活更加便捷、高效。

传统意义上的信息网络划分是以其传播载体的角度进行区分的,包括公用电话网、广播电视网和计算机网络。三网的形成和发展模式有其鲜明的历史特性,而在移动互联网技术飞速发展的今天,三网的边界也越来越模糊,其服务对象和服务功能相互交叉,互为补充,“三网融合”已是大势所趋,融合的边界也逐渐向广度、深度延伸,“IPv6技术”“物联网”“区块链”等的发展为信息网络的融合创新提供了新的思路和发展方向。

3.2.2 信息网络技术体系

克劳德·艾尔伍德·香农[①]在《通信的数学理论》中提出信息是可以被量化的,并指出通信的本质是数据交换。那么在信息网络中的各个节点之间是如何相互识别的?不同节点之间又是如何进行通信的?数据在信息网络中是以一种什么样的形式进行传输的?在本节中,将对信息网络的技术体系组成进行介绍,为深入理解信息网络奠定理论基础。

1. IP地址

在进行网络通信之前,首先要了解的就是IP地址,它是给每个连接在Internet上的主机(或路由器)分配的一个在全世界范围的唯一的标识符。例如在QQ上发送消息,消息中的信息是如何传送到对方的计算机中的呢?要在连接互联网的亿万台主机中找到某一台计算机,就需要知道代表那台计算机的唯一标识符,这就是IP地址。就如与人打电话,就必须知道对方的电话号码一样。IP地址就是每台主机在Internet中的“电话号码”,有了它就能收发信息(接打电话)了。

同电话号码一样,IP地址使用固定长度的数字来表示,现在常用的表示方法为IPv4地址,它是一个由32位二进制数组成的地址。人们在实际应用中为了便于表达,一般将这32比特数分为4段,每段8比特,然后将这4段8比特的二进制数转换为十进制数,十进制数之间用“.”隔开。这种方法叫作点分十进制表示法(Dotted Decimal Notation)。例如,地址

① 克劳德·艾尔伍德·香农,美国数学家、信息论的创始人,1940年在麻省理工学院获得硕士和博士学位,1941年进入贝尔实验室工作。香农提出了信息熵的概念,为信息论和数字通信奠定了基础。

10000000 01100100 00000011 00001010 用点分十进制表示法表示为 128.110.3.10。

IP 地址采用层次结构，由两部分构成，即网络号与主机号，网络号在前，主机号在后。其中，网络号用来标识主机所在的逻辑网络（类似于固定电话号码前的区号），主机号用来表示网络中的一个接口。一台 Internet 主机至少有一个 IP 地址，而且该 IP 地址是全球唯一的。如果一台 Internet 主机有两个或多个 IP 地址，则该主机属于两个或多个逻辑网络。

传统的 IP 地址编码方案采用所谓的“分类 IP 地址”，分别称为 A 类、B 类、C 类、D 类和 E 类。其中 A、B 和 C 类由全球性的地址管理组织在全球范围内统一分配，D 类和 E 类属于特殊地址。

IP 地址采用高位字节的高位来标识地址类别。IP 地址分类编码方案如图 3-2 所示。

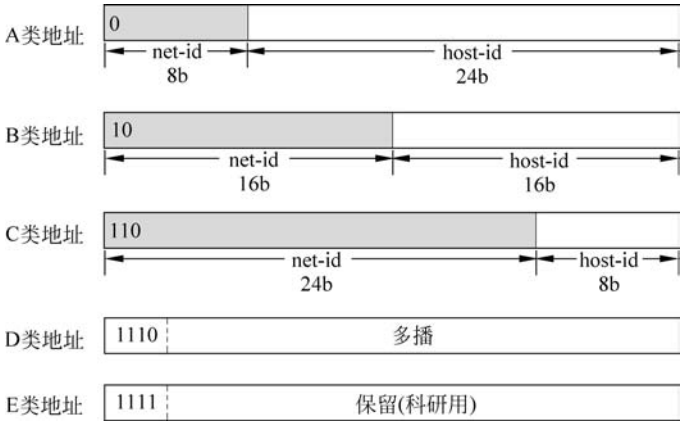


图 3-2　IP 地址分类编码方案

（1）A 类地址的第一位为 0，B 类地址的前 2 位为 10，C 类地址的前 3 位为 110，D 类地址的前 4 位为 1110，E 类地址的前 4 位为 1111。

（2）A 类、B 类和 C 类地址的网络号字段分别为 1 字节、2 字节和 3 字节长，A 类、B 类和 C 类地址的主机号字段分别为 3 字节、2 字节和 1 字节。

将 IP 地址划分为三个类别的原因是这样的：各种网络的差异很大，有的网络拥有很多主机，而有的网络上的主机很少。将 IP 地址划分为 A 类、B 类和 C 类可以更好地满足不同用户的要求。

当某个单位申请到一个 IP 地址时，实际上只是获得了一个网络号 net-id，具体的各个主机号 host-id 则由该单位自行分配，只要做到在该单位范围内无重复的主机号即可。

除了上述三类 IP 地址以外，还有两类使用较少的地址，即 D 类和 E 类地址。D 类地址是多播地址，E 类地址保留给以后使用。

A 类地址的 net-id 字段有 1 字节，由于最高位已经固定为 0，因此剩下的 7 位共能表示 126（2^7-2）个 A 类网络，这里减 2 的原因是：全 0 的 IP 地址是保留地址，意思是“本网络”；值为 127（即 01111111）保留作为本地软件环回测试（Loopback Test）本主机之用。后 3 字节是 host-id，每一个 A 类网络中的最大主机数量是 16 777 214（即 $2^{24}-2$）。减 2 的原因是：全 0 的 host-id 字段表示该 IP 地址是“本主机”所连接到的单个网络地址（例如，某一主机的 IP 地址是 126.100.10.8，则该主机所在的网络地址就是 126.0.0.0），而 host-id 为全 1 表示“所有的（all）”，因此全 1 的 host-id 字段表示该网络上的所有主机，即本网内广播。

整个A类地址空间共有2^{31}(即2 147 483 648)个地址,而IP地址全部的地址空间共有2^{32}(即4 294 967 296)个地址。可见A类地址占整个IP地址空间的50%。

B类地址的net-id字段有2字节,但前面2比特值已经固定(10),只剩下14比特可以变化,因此B类地址的网络数为16 384(2^{14})。请注意,这里不需要减2,因为这14比特加上最前面固定的2比特值10,无论如何也构不成全0或者全1。B类地址的每一个网络上的最大主机数是65 534(即$2^{16}-2$)。这里减2和A类网络一样,是因为要扣除全0和全1的主机号。整个B类地址空间共有1 073 741 824(2^{30})个地址,占整个IP地址空间的25%。

C类地址有3字节的net-id字段,最前面3比特的标识位是110,还有21比特可以变化,因此C类地址的网络总数是2 097 152(即2^{21},这里也不需要减2)。每一个C类地址的最大主机数是254(即$2^{8}-2$)。整个C类地址空间共有536 870 912(即2^{29})个地址,占整个地址空间的12.5%。

所有IP地址的使用范围如表3-1所示。

表3-1 所有IP地址的使用范围

网络类别	最大网络数	第一个可用的网络号	最后一个可用的网络号	每个网络中的最大主机数
A	126($2^{7}-2$)	1	126	16 777 214
B	16 384(2^{14})	128.0	191.255	65 534
C	2 097 152(2^{21})	192.0.0	223.255.255	254

一般不使用的特殊IP地址如表3-2所示。

表3-2 一般不使用的特殊IP地址

net-id	host-id	源地址使用	目的地址使用	代表的意思
0	0	可以	不可以	在本网络上的本主机
0	host-id	可以	不可以	在本网络上的某个主机
全1	全1	不可以	可以	只在本网络上进行广播(各路由器均不转发)
net-id	全1	不可以	可以	对net-id上的所有主机进行广播
127	任何数	可以	可以	用作本地软件环回测试

随着IP网络爆炸性地发展,更重要的是全球Internet的飞速发展,可用的IP地址空间正在缩小,核心的Internet路由器处理能力也逐渐耗尽。Internet面临着必须尽早解决的问题,这就是:

(1) IPv4网络地址的耗尽问题。

(2) 由于Internet的发展,Internet的路由选择表的大小在迅速、大量的增加。随着更多的C类地址加入Internet,新网络信息的大量充斥威胁到Internet路由器的处理能力。

在IPv4地址结构下,A类和B类地址构成了75%的IPv4地址空间,但只有少数公司和组织能够分配到一个A类或B类网络号。C类网络号比A类和B类网络号要多得多,但它们仅仅占了可能的40亿(2^{32})IP地址的12.5%,各类地址所占比例如图3-3所示。

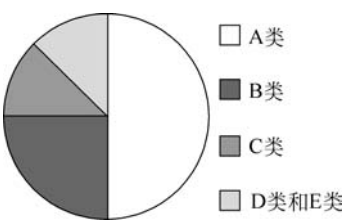


图3-3 各类地址所占比例

2019年11月25日,欧洲地区互联网注册网络协调中心(RIPE NCC)宣布,其最后的IPv4地址空间储备池已完

全耗尽，所有 43 亿个 IPv4 地址已分配完毕。

人们一直在寻求解决 IPv4 地址危机的办法，常用的方法有：

(1) 无类域间路由(CIDR)和可变长子网掩码(VLSM)。

(2) 私有 IP 地址[①]与网络地址转换(Network Address Translation，NAT)。

人们在 A 类、B 类和 C 类地址段中各取了一部分地址空间作为私有地址。这部分规划保留的地址是：A 类 IP 地址中的 10.0.0.0～10.255.255.255；B 类 IP 地址中的 172.16.0.0～172.31.255.255；C 类 IP 地址中的 192.168.0.0～192.168.255.255。

私有地址不能直接接入 Internet，也不会被 Internet 路由。使用了私有 IP 地址的本地网络中的计算机如果需要连接 Internet，需要借助于专门的技术，即 NAT。

NAT 允许一个整体的本地网络在其内部均使用私有 IP 地址，在 Internet 上只使用一个或少量的公用 IP 地址。当内部节点需要与外部网络进行通信时，NAT 可将内部私有 IP 地址翻译成外部公有 IP 地址，从而得以正常访问 Internet。这样一来，就可以使用较少的公有 IP 地址，解决更多内部节点机器的 Internet 访问问题，从而有效地缓解了 IP 地址不足的问题。

(3) IPv6(彻底的根本解决方法)。

IPv6 把原来的 IPv4 地址增大到了 128 位，其地址空间大约是 3.4×10^{38}，是原来 IPv4 地址空间的 2^{96} 倍，这样就可以彻底解决地址不足的问题。另外，IPv6 并没有完全抛弃原来的 IPv4，并且在若干年内都会与 IPv4 共存。IPv6 使用一系列固定格式的扩展首部取代了 IPv4 中可变长度的选项字段。IPv6 对 IP 数据报协议单元的头部进行了简化，仅仅包含 7 个字段(IPv4 有 13 个)。这样，当数据报文经过中间的各个路由器时，各个路由器对其处理的速度可以更快，从而可以提高网络吞吐率。IPv6 内置了支持安全选项的扩展功能，如身份验证、数据完整性和数据机密性等。

2. 域名系统服务

在实际要定位一台主机时，往往使用的是主机的域名而非 IP 地址，这个域名就好比人名一样比较容易记忆，而网络中的域名服务器用来关联每个域名与 IP 地址，使用域名服务进行联系的机制就是域名系统(Domain Name System，DNS)。DNS 是 Internet 使用的命名系统，用来把便于人们使用的机器名字转换为 IP 地址。DNS 其实就是名字系统。

用户与 Internet 上某个主机通信时，必须知道对方的 IP 地址。然而用户很难记住长达 32 位的二进制主机地址。即使是点分十进制表示的 IP 地址也并不太容易记忆。但在应用层为了方便用户记忆各种网络应用，更多的是使用主机名字。那为什么机器在处理 IP 数据报时要使用 IP 地址而不使用域名呢？这是因为 IP 地址的长度是固定的 32 位(IPv6 地址是 128 位)，而域名的长度是不固定的，机器处理起来比较困难。

Internet 的 DNS 被设计成为一个联机分布式数据库系统，并采用客户-服务器方式。DNS 大多数名字都在本地进行解析，仅少量解析需要在 Internet 上通信，因此 DNS 的效率很高。由于 DNS 是分布式系统，即使单个计算机出了故障，也不会妨碍整个 DNS 的运行。

---

① 所谓私有地址就是在 A、B、C 三类 IP 地址中保留下来为内部网络分配地址时所使用的 IP 地址。私有地址主要用于在局域网中进行分配，在 Internet 上是无效的。这样可以很好地隔离局域网和 Internet。私有地址在公网上是不能被识别的，必须通过 NAT 将内部 IP 地址转换为公网上可用的 IP 地址，从而实现内部 IP 地址与外部公网的通信。

域名到IP地址的解析过程的要点如下：当某一个应用进程需要把主机名解析为IP地址时，该应用进程就调用解析程序，并成为DNS的一个客户，把待解析域名放在DNS请求报文中，以UDP(用户数据报)方式发给本地域名服务器(使用UDP是为了减少开销)。本地域名服务器在查找域名后，把对应的IP地址放在回答报文中返回。应用进程获得目的主机的IP地址后即可进行通信。

若本地域名服务器不能回答该请求，则此域名服务器就暂时成为DNS中的另一个客户，并向其他域名服务器发出查询请求。这种过程指导能够回答该请求的域名服务器为止。

原来的顶级域名共分为三大类。

(1) 国家顶级域名nTLD：采用ISO 3166规定。如：cn表示中国，us表示美国，uk表示英国，等等。国家顶级域名又常记为ccTLD(cc代表国家代码)。

(2) 通用顶级域名gTLD：最先确定的通用顶级域名有7个，即com(公司企业)、net(网络服务机构)、org(非营利性组织)、int(国际组织)、edu(美国专用的教育机构)、gov(美国的政府部门)、mil(美国的军事部门)。

截止到2011年初，又陆续增加了13个通用顶级域名：aero(航空运输企业)、asia(亚太地区)、biz(公司和企业)、cat(使用加泰隆人的语言和文化团体)、coop(合作团体)、info(各种情况)、jobs(人力资源管理者)、mobi(移动产品与服务的用户和提供者)、museum(博物馆)、name(个人)、pro(有证书的专业人员)、tel(Telnic股份有限公司)、travel(旅游业)。

(3) 基础结构域名(infrastructure domain)：这种顶级域名只有一个，即arpa，用于反向域名解析，因此又称为反向域名。

值得特别注意的是，2011年6月20日在新加坡会议上正式批准新顶级域名(New gLTD)，因此任何公司、机构都有权向ICANN申请新的顶级域名。新顶级域名的后缀特点，使企业域名具有了显著的、强烈的标志特征。因此，新顶级域名被认为是真正的企业网络商标。

在国家顶级域名下注册的二级域名均由该国家自行规定。例如，就顶级域名jp的日本而言，其将教育和企业机构的二级域名定义为ac和co，而不是edu和com。而我国把二级域名划分为“类别域名”和“行政区域名”两大类。

我国的类别域名共7个，分别为ac(科研机构)、com(工、商、金融等企业)、edu(教育机构)、gov(政府机构)、mil(国防机构)、net(提供互联网络服务的机构)、org(非营利性组织)。

我国的行政区域名一共34个，适用于我国的各省、自治区和直辖市。

我国修订的域名体系允许直接在cn的顶级域名下注册二级域名。这显然给我国的Internet用户提供了极大的方便。关于我国的互联网络发展现状以及各种规定(包括申请域名的手续)，均可在中国互联网络信息中心(CNNIC)的网址上找到。

3. 网络协议与TCP/IP

有了IP地址和域名服务之后，那么处于网络上的两台主机之间是如何进行网络通信的呢？如何发送和接收信息？如何进行状态反馈？这些都需要在统一规则的约束下进行，而这个规则就是网络协议。

计算机网络中大量的数据在进行交互，这些数据的交互是在一定的规则、标准或约定下进行的，这些规则明确规定了所交换的数据的格式以及有关的同步问题。这些为网络中的数据交换而建立的规则、标准或约定称为网络协议(Network Protocol)，简称为协议。网络

协议对于计算机网络是至关重要的，它的存在使网络上各种设备能够相互交换信息。网络协议主要由以下三个要素组成。

（1）语法，规定了数据与控制信息的结构或格式，包括数据出现的顺序。

（2）语义，规定了各种控制信息的意义，说明通信双方该怎么做。

（3）时序，也称为同步，规定了事件实现的顺序。

简单来说，就像中国和法国的两家企业的领导一起开会，语法就是大家都能理解的语言的语法（假定这种语言是英语）；语义就是使用的英语单词和语句的意思；时序就是两位领导的秘书事先商量好谁先说、谁后说，先讨论什么内容、后讨论什么内容，语速是快还是慢等。

由此可见，网络协议是计算机网络不可缺少的组成部分。协议通常有两种不同的形式：一种是使用便于人阅读和理解的文字描述；另一种是使用让计算机能够理解的程序代码。这两种不同形式的协议，都必须能够对网络上信息交换过程做出精确地解释。

目前，在 Internet 以及众多的局域网中使用的网络协议体系结构都是 TCP/IP 模型。TCP/IP 体系结构如表 3-3 所示。

表 3-3　TCP/IP 体系结构

层　次	功 能 描 述
应用层	定义了 TCP/IP 及主机应用程序与网络运输层服务之间的接口
运输层	提供主机之间的通信会话管理，定义传输数据时的服务级别和连接状态
网络层	将数据装入 IP 数据报，包括用于在主机间及经过网络转发数据报时所用的源地址和目标地址信息，实现 IP 数据报的路由和寻址
网络接口层	通过网络，实现数据的实际物理传输，包括直接与传输介质接触的硬件设备、如何将比特流转换为电信号等

TCP/IP 体系结构的核心为网络层的 IP 与运输层的 TCP，具体如下。

IP：IP 是 Internet Protocol 的缩写，中文名称为网际互联协议，它是 TCP/IP 体系结构中的网络层协议。IP 可以提高网络的可扩展性：一是解决网络互联问题，实现大规模、异构网络的互联互通；二是分割顶层网络应用和底层网络技术之间的耦合关系，以利于两者的独立发展。需要注意的是，IP 只为主机提供一种无连接、不可靠的、尽力而为的数据包传输服务。

TCP：TCP 是 Transmission Control Protocol 的缩写，即传输控制协议，它是一种面向连接的、可靠的、基于字节流的传输层通信协议。互联网络与单个网络有很大的不同，因为互联网络的不同部分可能有截然不同的拓扑结构、带宽、延迟、数据包大小和其他参数。TCP 的设计目标是能够动态地适应互联网络的这些特性，而且具备面对各种故障时的健壮性。

3.2.3　信息网络组成结构

在 3.2.2 节中，围绕信息网络的技术体系进行了解读和学习，那么真实的信息网络是如何搭建起来的呢？网络节点是如何接入的呢？信息网络的拓扑结构有哪些类型呢？本节将为读者一一解读信息网络的组成结构。

1. 信息网络接入技术

截至2021年12月底,我国固定宽带家庭普及率超过90%,固定宽带用户达5.36亿户,其中光纤接入用户达5.06亿户,占固定宽带用户的比重达94.3%,远超OECD(经济合作与发展组织)国家26.8%的平均水平,仅次于新加坡(99.7%),位居全球第二。移动宽带普及率远超预期目标,移动电话用户数量达16.43亿户。其中5G移动电话用户达3.55亿户,远超全球平均水平。网络家庭普及率的提升离不开网络接入技术的发展,从接入技术上可以分为数字用户线接入、光纤接入、光纤同轴混合网接入、局域网接入、无线接入等。

1) 数字用户线接入

数字用户线接入是一种通过普通的电话线路实现网络接入,能够支持电话和网络接入的服务模式。其中ADSL(Asymmetrical Digital Subscriber Line,非对称数字用户线)能够同时支持电话和网络服务,素有"网络快车"之美誉,其传输距离取决于数据率和用户线的线径(线径越细,衰减越大,传输距离越短)。ADSL在用户线的两端各安装一个ADSL调制解调器,采用自适应调制技术使用户线能够传送尽可能高的数据率。ADSL的上行信道带宽低于下行信道带宽。

2) 光纤接入

光纤接入是通过光纤直接连接到用户终端的网络应用,把要传送的数据由电信号转换为光信号进行通信,在光纤的两端分别都装有"光猫"进行信号转换,具有通信容量大、质量高、性能稳定、防电磁干扰、保密性强等优点。在光纤通信中,光纤扮演着重要角色。在接入网中,光纤接入也是发展的重点。光纤接入方式可分为如下几种:FTTB(Fiber To The Building,光纤到大楼)、FTTC(Fiber To The Curb,光纤到路边)、FTTZ(Fiber To The Zone,光纤到小区)、FTTF(Fiber To The Floor,光纤到楼层)和FTTH(Fiber To The Home,光纤入户)等。

3) 光纤同轴混合网接入

光纤同轴混合网(Hybrid Fiber Coaxial,HFC)是在目前覆盖面很广的有线电视网络基础上开发的一种结合光纤与同轴电缆的宽带接入网,采用频分复用技术,除可传送电视节目外,还能提供电话、数据和其他宽带交互型业务,扩展性较好。

4) 局域网接入

局域网接入是将一个局域网连接到Internet,现在常用的方法是通过路由器将局域网与Internet连接起来。

5) 无线接入

无线接入技术(Wireless Access Technology)是通过无线介质将用户终端与网络节点连接起来以实现信息传递的一种技术。它与有线接入最重要的区别是可以向终端提供移动接入服务。从终端接入类型来分,无线接入可以分为集群移动无线接入、蜂窝移动网络、卫星通信网络等。生活中常用的接入方式包括蓝牙(Bluetooth)、CDMA2000、GSM、5G、Wi-Fi(Wireless Fidelity)、射频(RF)等。

2. 信息网络传输介质

传输介质是网络中连接收发双方的物理通路,也是通信中实现信息传送的载体。传输介质通常分为有线传输介质(导向型介质)和无线传输介质(非导向型介质)。

1）有线传输介质

（1）双绞线。

双绞线由两根分别包有绝缘材料的铜线螺旋状地绞合在一起，芯线为软铜线，线径为0.4～1.4mm。两线绞合的目的是减少相邻线对之间的电磁干扰，通信距离一般为几到十几千米。距离太长时需要加放大器以便将衰减了的信号放大到合适的数值（对于模拟传输），或者加上中继器以便将失真了的数字信号进行整形（对于数字传输）。由于双绞线价格便宜且性能也不错，因此使用非常广泛。

（2）同轴电缆。

同轴电缆（Coaxial Cable）由一根内导体铜质芯线外加绝缘层、密集网状编织导电金属屏蔽层以及外包装保护材料组成，其特点是高带宽及良好的噪声抑制性。同轴电缆的带宽取决于电缆长度，1km 的电缆可以达到 1～2Gb/s 的数据传输速率。

（3）光纤与光缆。

光纤通信就是利用光导纤维（简称光纤）传递脉冲光来进行通信。由于可见光的频率非常高，约为 10^8MHz 的量级，因此一个光纤通信系统的传输带宽远远大于目前其他各种传输媒介的带宽。此外，光纤还具有传输损耗小、中继距离长、抗雷电和电磁干扰、性能好、无串音干扰、保密性好、体积小、重量轻等特点。

2）无线传输介质

对于有线传输介质来讲，若是通信线路要通过一些高山或岛屿，有时是很难施工的。即使在城市中，敷设电缆也不是一件很容易的事情。当通信距离很远时，敷设电缆既昂贵又费时。但利用无线电波在自由空间的传播就可以较快地实现多种通信。因此，就将自由空间称为无线传输介质（非导向型传输媒体）。无线传输介质包括：

（1）短波。

短波通信主要靠电离层的反射。通信频率范围为 3～30MHz，通常称为高频（HF）段。由于电离层随季节、昼夜以及太阳黑子活动而变化，因此通信质量并不稳定。

（2）微波。

无线电微波通信在数据通信中占有重要地位。微波的频率范围为 300MHz～300GHz（波长为 1～10cm），但主要是使用 2～40GHz 的频率范围。微波在空间中主要是直线传播，并且能够穿透电离层进入宇宙空间，因此它不像短波那样可以经电离层反射传播到地面上很远的地方。由于地球表面是一个曲面，因此一般在山顶建立微波中继站（简称“微波站”）。微波站的通信距离一般为 30～50km，当微波天线高达 100m 时，通信距离可以达到 100km。为实现远距离通信，必须在一条微波通信信道的两个终端之间建立若干个中继站。中继站把前一站送来的信号经过放大后再发送到下一站，故称为“接力”。

（3）卫星。

常用的卫星通信方法是利用位于约 36 000km 高空的人造地球同步卫星作为中继器的一种特殊形式的微波接力通信。和微波接力通信类似，卫星通信的频带很宽，通信容量很大，信号所受到的干扰也较小，通信比较稳定，并且卫星通信的通信费用与通信距离无关。

（4）红外线通信和激光通信。

红外线通信和激光通信就是把要传输的信号分别转换为红外光信号和激光信号直接在

自由空间沿直线进行传播,它比微波通信具有更强的方向性,难以窃听、插入数据和进行干扰,但红外线和激光对雨雾等环境干扰特别敏感。

3. 网络数据交换技术

网络的主要目的是实现网络节点间的数据传输和信息交换,三种典型的网络拓扑结构如图 3-4 所示,其中各中心节点承担着数据传输和交换的角色。常见的数据交换技术有电路交换技术、报文交换技术和分组交换技术,基于不同交换技术又可将网络划分为电路交换网络、报文交换网络和分组交换网络。

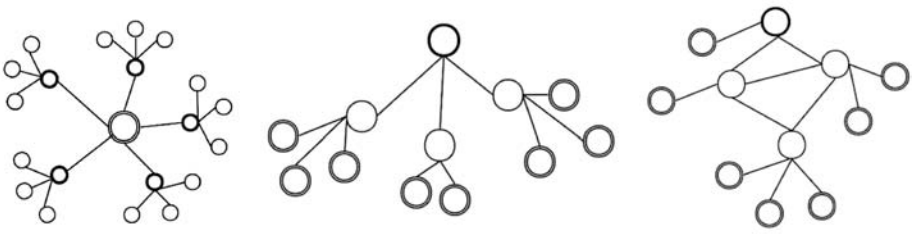


图 3-4 三种典型的网络拓扑结构

电路交换技术采用通信链路资源独占模式,进行通信传输前,节点之间必须先建立一条专用的物理通信路径,在整个传输期间线路一直被独占,直到通信结束后才被释放。其中,电话网络是最典型的电路交换网络。

报文交换技术以报文为数据交换单位,报文中携带有目标地址、源地址、报文内容等信息,其运行原理类似于邮件投递。在进行通信前,双方不需要预先建立专用通信线路,交换节点存储接收到的报文信息,根据报文信息判断其目标地址后选择空闲线路进行路由转发,也支持报文的广播传输。报文交换由于数据在交换节点需要经历存储转发,会造成通信时延,很少被应用于目前的网络数据交换。

分组交换技术是目前应用范围最广的一种交换技术,采用存储转发技术,通过将报文拆分成多个分组进行传送,每个分组的长度均有一个上限,从而降低分组缓冲区的大小,分组中均携带有源节点、目标节点地址信息。与报文交换技术所不同的是,分组交换技术通过将大报文拆分成小的分组后在交换节点进行存储转发,各分组独立地选择传输路径进行并行转发,缩短了整体的数据传输时延。同时由于分组较小,传输出错率和重发数据量大小减少,提高了传输的可靠性。相对比以上三种主流的数据交换技术,报文交换技术和分组交换技术在对网络信道利用率上要优于电路交换技术,其中分组交换技术在网络传输时延上要比报文交换技术小,适合于交互式通信场景。

4. 无线通信技术概述

如今,无线通信技术已经经历了数代的发展,第一代通信技术标志着个人移动通信的诞生,其中“大哥大”使用的就是 1G。1G 采用模拟通信技术,是最初的模拟,仅限语音的蜂窝电话标准,表示传递信息所使用的电信号或电磁波信号往往是对信息本身的直接模拟,例如语音(电话)、静态图像(传真)、动态图像(电视、可视电话)等信息的传递,其中用户的语音信息的传输是以模拟语音方式出现的。

此时，人们需求的不断提高，使人们意识到个人移动通信的必要性和可能性，同时创造了个人移动通信的原始产业链，产生了设备供应商、移动通信运营商，但由于技术的原因，通信存在质量差、体积大、缺乏规模效应且价格昂贵的缺点。这也推动了2G时代的诞生，这标志着个人移动通信进入成长期，在这个阶段，通信质量得到了极大的提高，几乎完全满足消费者的需求。

由于通信技术的快速成长，移动通信不再是一个孤立的系统而是开放性的，即进入3G时代。此时，各个国家和地区的移动通信网将融合为一个整体，除了传统的语音业务外，更多的是数据和多媒体业务。

4G时代象征着移动通信的发展进入成熟期，是基于IP的高速移动通信网络，是移动通信技术发展史上的一次重大变革。它比3G的传输容量大，速率更快，并且具备了长期演进语音承载(VoLTE)通信技术，实现了系统向宽带无线化和无线宽带化的演进。

随着5G时代的到来，面向个人和行业的移动应用快速发展，移动通信相关产业生态将逐渐发生变化。5G不仅仅是高速率、宽带宽、高可靠、低时延的无线接入技术，而且是面向用户体验和业务应用的智能网络。同时，5G技术充分利用了物联网，整体来说提高了经济效益和社会效益，很大程度上促进了互联网技术的持续性发展。截至2019年12月，我国已经建成5G基站超过13万个，5G产业链推动人工智能与物联网结合发展到智联网。

在信息网络飞速发展的今天，地球上仍然有超过30亿人口，约70%的地理空间因人口密度大、地理环境差、建设成本高昂等原因未能享受到移动互联网的覆盖。早在2015年，SpaceX首席执行官埃隆·马斯克(Elon Musk)宣布“星链卫星互联网服务”项目，旨在为全世界用户提供高速互联网接入、特别是农村和偏远地区。

按照SpaceX此前的计划，Starlink未来计划发射12 000颗卫星组建覆盖在地球周围的通信网络体系，首批1600颗卫星部署位于1150km的轨道高度，其中800颗卫星用于覆盖北美地区；第二批发射2825颗卫星分为4组，分布部署于不同高度的轨道之上，完成全球组网；第三批在340km的更低轨道上发射7518颗卫星。截至2022年3月底，累计超过2000颗卫星部署于地球轨道之上。根据公开的网络测试显示，Starlink的测试版速率已突破160Mb/s，超过美国95%的宽带连接。相比于常规基站布局通信技术，星链互联网不受地面基础设施限制，满足全世界各个角落，无论偏远山区、高原、海底的全天候、低成本接入。

SpaceX并不是第一个利用卫星互联网提供通信网络服务的公司，早在20世纪80年代，美国摩托罗拉公司就启动了“铱星计划”，在围绕地球近地轨道之上建立分布均匀的卫星协作体系，它的天上部分是运行在7条轨道上的卫星，每条轨道上均匀地分布着11颗卫星，组成一个完整的星座。它们就像铱原子核外的77颗电子围绕其运转一样，因此被称作铱卫星。后来经过计算证实6条轨道就可满足建设需求，于是卫星总数减为66颗，但仍习惯称作铱卫星。“铱星计划”总投资34亿美元，于1996年开始试验发射，1998年投入运行。然而在“铱星计划”投入运行后未能有效占领市场，公司亏损严重，摩托罗拉公司不得不将铱星公司申请破产保护，从投入运行到终止不到短短半年时间，最终于1999年3月17日对外停止服务。“铱星计划”的失败并没有阻止人类继续对卫星互联网的探索和不懈追求，越来越多的国家和企业投入卫星互联网领域。

早在2017年底到2018年之间，我国就发布了多个通信卫星星座项目，由中国航天科技集团和航天科工集团发起的“鸿雁”“虹云”“行云”计划等开启我国在卫星互联网领域的探索

和实践。根据2020年公布的“新基建”范围划定,国家发改委首次将低轨卫星互联网建设纳入“新基建”的建设范围,我国低轨卫星互联网发展将迎来重大发展机遇,截止到2022年在轨低轨卫星规模达800余颗。在卫星互联网不断探索的同时,基于太赫兹(THz)的6G通信技术也是各国火热争夺的焦点。6G卫星互联网不是简单的卫星互联网,而是“空天地”一体化网络,包括卫星与地面、卫星与卫星、卫星与高空之间的通信。

2020年11月6日,我国在太原卫星发射中心成功利用长征六号运载火箭将电子科技大学和国星宇航等单位联合研制的全球首颗6G试验卫星“电子科技大学号”送入太空,开创了国内6G卫星互联网探索新时代。其中太赫兹通信技术是6G的关键技术之一,具有高频率、窄波束、强穿透力等特性,可解决带宽受限与可靠传输的问题,适合星地通信、保密通信、空间通信、军事通信等的特殊要求。此次“电子科技大学号”的成功发射是国内太赫兹通信在空间应用场景下的首次技术验证。相比5G网络传输,6G传输速度将达到100Gb/s~1Tb/s,通信时延缩短到0.1ms以内,采用太赫兹频段通信,网络通道容量将大幅提升,而且它的发展趋势是将卫星通信与地面通信网络相融合,从而真正实现全球无盲区通信,即便是在荒无人烟的沙漠和茫茫大海上也能实现全天候、无障碍、即时网络通信。

3.2.4 信息网络运行机制

在前面的章节中介绍了信息网络技术体系和网络组成结构,那么信息网络是如何进行数据传输和共享交换的?信息又是如何准确地传输到信息需求方?在传输过程中又会经历哪些环节?在本节中,将向读者完整介绍信息网络背后的运行机制,帮助读者更加深刻地掌握信息网络的原理。

邮政EMS快递面单如图3-5所示,与传统的邮政投递系统相类似,网络的数据传输同样需要具备寄件人、收件人、中转中心、邮寄对象、唯一邮件编码等基础元素。结合前面章节所学习的内容,目的地IP、源地址IP代表网络中的每一个节点,唯一性的特征保证了信息在寄件人与收件人之间的准确投递。信息传输协议定义了跨网络间信息的传输格式和交换标准,保证了数据的准确传输。数据部分代表网络中所交换的对象和内容。以TCP传输为例,信息传输过程中,信息包由首部和正文数据两部分构成,在发送端发送信息时,在经过的每一层传输协议都会对信息包进行封装,而在信息包到达接收端时,接收系统按照TCP层级进行逐层解封从而获取原始信息包,TCP信息首部结构如图3-6所示。

图3-5 邮政EMS快递面单

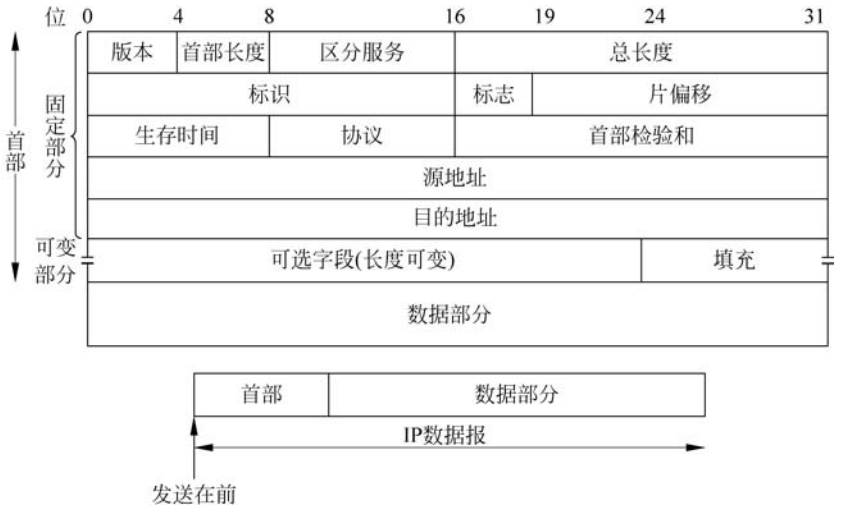


图 3-6　TCP 信息首部结构

以 Web 信息浏览服务为例，它不是普通意义上的物理网络，而是一张附着在 Internet 上的覆盖全球的“信息网”，是一个大规模的、联机式的信息储藏所。严格来讲，Web 是一个技术系统，使用链接的方法能非常方便地从 Internet 上的一个站点访问另一个站点（也就是所谓的“链接到另一个站点”）。Web 信息服务结构如图 3-7 所示。提供共享信息资源的站点称为“Web 网站”；承载资源信息内容的服务器称为“Web 服务器”。Web 服务器、超文本传输协议（HTTP）、浏览器是构成 Web 的三个要素。

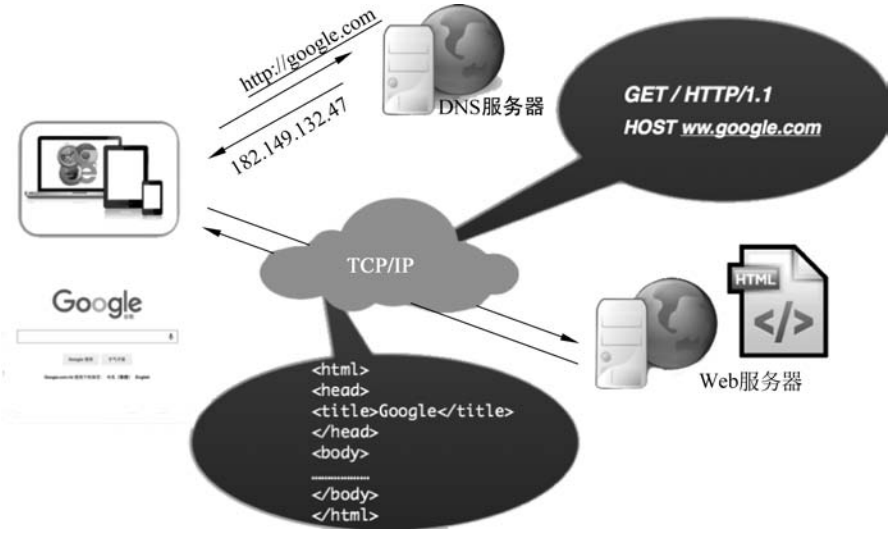


图 3-7　Web 信息服务结构

通过统一资源定位器（Uniform Resource Locator，URL），标识和寻址分布在整个 Internet 上的信息资源。Web 中信息资源是巨大的，每个承载着信息内容的网页都必须具有一个唯一的名称标识，通常称为 URL 地址，俗称“网址”，否则信息再丰富也不能实现便捷地访问浏览。为保证信息资源命名的唯一性，URL 制定了统一的格式和规则。

URL的一般使用格式如下为 scheme://host:port/path/filename。

- scheme：通信协议，指示该信息资源服务的协议类型。URL中通信协议名称如表3-4所示。如果为HTTP服务，可省略“http://”。

表3-4 URL中通信协议名称

协议名称	功能
file	本地计算机上的文件资源
ftp	通过FTP访问的信息资源
gopher	通过Gopher协议访问的信息资源
http	通过HTTP访问的信息资源
https	通过安全的HTTP访问的信息资源
mailto	资源为电子邮件地址，通过SMTP访问
news	通过NNTP访问的信息资源

- host：主机名，只是提供信息服务的服务器域名或IP地址。
- port：端口号，为可选项，只是提供信息服务所使用的端口号。如果使用的是Internet上信息服务的默认端口号，此项可以省略。例如，HTTP服务的默认端口号为80，如不重新设置改变端口，则端口号80就可以省略。
- path：路径。指示资源文件在服务器中存放的路径。
- filename：文件名。指示要访问的存放在服务器中指定路径下的资源文件的文件名。如果要访问的资源文件为网站的主页，则一般可省略此项。

3.3 大数据时代下的信息网络

大数据时代，越来越多的新技术得以与信息网络相结合，以极快的发展速度、创新的服务模式融入人们的生活中，并且在各个领域发挥着不可替代的作用，信息网络已成为数据时代中不可或缺的技术之一。新时代背景下的信息网络将与新技术不断融合，并对社会发展产生更多、更新的影响。

3.3.1 信息网络与新技术的融合

大数据时代的到来，信息网络范围变得更加广泛，信息数据的交互渠道、方式等更加多样化，为了满足人们更高、更加多元化的需求，产生了更多先进的信息网络技术，这些技术的产生很好地解决了信息网络中出现的一系列问题，为人们的生活提供了更大的便捷。在本节中将对大数据时代背景下的新型技术进行详细介绍。

1. 人工智能

大数据技术的飞速发展，为人类社会的工作和生活带来了极大的便利，但是同时也带来了比以往信息时代中更为复杂的网络安全问题，然而传统的信息网络技术在识别和判断网络安全问题时，通常不会对所得数据信息的真实性和准确性进行明确判定，在广泛收取数据的同时，无法精准判定网络数据中的虚假信息、攻击性数据等。人工智能技术能够通过模拟人的思维，进行大数据时代海量信息数据的精准处理，从而屏蔽虚假、有攻击性的数据信息，

通过有针对性、智能化的处理技术，有效缩短现代网络技术中对各类型信息的处理效率，提高网络管理的水平。

人工智能的优势在于具有处理模糊信息能力和协作能力，具备学习能力和处理非线性问题的能力且计算成本小等特点。它的具体应用可以包括以下几个方面。

(1) 网络安全方面。人工智能在计算机网络技术中网络安全方面的应用是可以智能分辨并自动处理垃圾软件。当邮件收到垃圾邮件时，人工智能可以自动识别和筛选，并做出一系列的措施，如自动标识垃圾邮件、含有敏感词时可以检测邮件并且阻止进入，可以最大程度地保护用户的安全。目前很多邮箱如网易邮箱、新浪邮箱等都已经使用了人工智能技术，并取得了很好的效果。人工智能可以进行信息识别，模糊处理存在的不确定信息，一旦有危险信息存在，有可能进入网络系统，防火墙就可以自动识别并自发地排除，保障网络的安全。这种人工智能技术在很多杀毒软件中都有使用，如金山毒霸、360 安全卫士等。

(2) Agent 技术的使用。人工智能 Agent 技术的出现从整体上提升了网络安全防护水平。人工智能 Agent 技术的出现，一方面通过对知识库、推力器等实体元件的使用，增强了网络安全防护过程中对复杂问题的存储、处理能力；另一方面该技术能够加强网络安全系统对周围环境的判断，当单个处理器处于工作状态时，可以智能调用通信网络对整体系统进行功能调用，共同完成任务。此外，该项技术还能够根据网络安全管理系统中的用户需求进行定制化服务。人工智能根据用户需求相关信息，智能筛选用户需要的信息并定位。该技术的出现大大提升了用户对网络安全信息筛选的效率，同时个性化、定制化的服务也为用户使用网络安全软件提供了便捷。用户在使用网络进行有用信息的选择过程中，系统将大量虚假、具有潜在攻击威胁的信息进行过滤，用户所得到的信息在绝大多数情况下是真实的、有效的。

(3) 网络管理方面。人工智能还能够依托专家知识库等技术进一步提升计算机网络的安全性。随着现代信息技术的飞速发展，信息网络与人工智能的结合过程中实现了质的转变，表现出显著的动态性特征。这些变化一方面为网络管理提供了巨大的便捷，另一方面也给网络管理工作带来了更大的难度。利用知识库系统能够更专业地解决系统网络管理问题，提升网络管理水平和工作效率。

(4) 家居方面。社会的发展和人们生活水平的日益提高使人们对信息网络的要求更高。人们对居家也提出了更高的要求，人工智能应用的可以更好地提高人们的生活质量。例如，可以控制门窗的闭合；可以随时操控来调整居家环境，让室内环境更加舒心等。因此，未来智能家居的应用范围会更加广泛，人们可以享受到更加优质的服务和生活质量。

此外，人工智能也被应用到农业生产、军事、医药等领域中，极大地推动着社会的发展。火车、高铁、地铁、飞机等交通工具的发展为人们的出行提供了极大的便利，随着出行人流量的增加，传统的人工检票可能会导致乘客排队或者出现漏检的可能。人脸识别系统的应用能够节约乘客的时间，极大地提高了效率。除了效率高之外，运用人脸识别系统后乘客安全性也大大提高。通过采集身份证照片和摄像头抓拍照片，利用人脸识别技术将人和证件进行匹配，同时将票面二维码信息和身份证件信息进行比对，根据检票系统业务定义规则，完成票、证、人自动验证检票功能。犯罪嫌疑人一旦“刷脸”，就会被迅速识别，并发出报警信息……这种人脸识别系统的应用大大提升了安检效率，有效提升了乘客的安全感。

除了交通工具使用人脸识别技术外，为了提高破案率，快速抓捕犯人，保障百姓安全，人

脸识别技术在安防领域也得到了广泛的运用。我们曾经会惊奇于警方能够捕捉摄像头中的影像来确认谁是犯罪嫌疑人,也感叹过系统能从海量级的人物照片库中找到被通缉人员,这些功能都是通过人脸识别技术实现的。“刷脸支付”越来越多地被广大消费者接受,也越来越普遍。刷脸支付能够保证消费者的资产安全,而且实现简单,通常情况下只需将资产账户信息绑定支付平台,并将支付平台绑定人脸识别业务,再提交近期照片和个人身份信息完成审核即可。当消费者刷脸支付时,系统会根据捕捉到的人脸与数据库中已被提交的照片信息进行对比,如果吻合,支付平台将自动连接消费者资金账户,在消费者确认过支付金额后,即可进行支付。人脸识别支付在保障支付安全的同时,能够实现流程短、耗时少的快捷支付。

受益于零售行业的数字化转型,人工智能渗透到零售各个价值链环节,实现了消费场景流程的全覆盖。在新零售的商业图谱下,人工智能助力零售商强化与消费者的互动并提供个性化商品和服务;同时,通过消费者数据优化货架布局,提升坪效以节约成本,提高消费者的消费体验。消费场景流程全覆盖具体为:消费者进店时先对消费者进行人脸识别,以获取消费者的基本信息,如是否是会员,其历史消费、购买力、偏好等信息;在消费者购物过程中,对商品状态进行监测,如压力感应、图像识别等,这样可以使零售门店综合消费者数据,以保证店内商品布局的最优决策。此外通过对商品的流动速度和库存的统计,以保证店内备货成本维持较低水平。在消费者离店后,对消费者行为进行分析,如通过行为、情绪识别或轨迹跟踪等。零售门店可以利用人工智能实现精准营销,如线上进行 APP 智能推荐,线下进行商品种类优化和位置调整等。

2. 数字货币

数字货币是电子货币形式的替代货币。数字货币是一种不受管制的、数字化的货币,通常由开发者发行和管理,被特定虚拟社区的成员所接受和使用。欧洲银行业管理局将虚拟货币定义为价值的数字化表示,不由央行或当局发行,也不与法定货币挂钩,但由于被公众所接受,所以可作为支付手段,也可以电子形式转移、存储或交易。数字货币的核心特征主要体现在三个方面:①由于来自某些开放的算法,数字货币没有发行主体,因此没有任何人或机构能够控制它的发行;②由于算法解的数量确定,因此数字货币的总量固定,这从根本上消除了虚拟货币滥发导致通货膨胀的可能;③由于交易过程需要网络中的各个节点的认可,因此数字货币的交易过程足够安全。数字货币的典型代表为比特币和莱特币。

根据麦肯锡的测算,从全球范围看,区块链技术在 B2B 跨境支付与结算业务中的应用大大降低了每笔交易成本,即区块链应用可以帮助跨境支付与结算业务交易参与方节省约 40%的交易成本,其中约 30%为中转银行的支付网络维护费用,10%为合规、差错调查以及外汇汇兑成本。未来,利用数字货币和区块链技术打造的点对点支付方式将省去第三方金融机构的中间环节,不但 24 小时实时支付、实时到账、无隐性成本,也有助于降低跨境电商资金风险及满足跨境电商对支付清算服务的及时性、便捷性需求。

低成本的资金转移和小额支付越来越受到使用者好评。电子支付使流通中现金在货币总量中的比重不断下降,最明显的就是银行业金融机构中发生的电子支付业务大幅增加,此外,在第三方支付方面,非银行支付机构累计发生网络支付业务也在逐年增加。电子支付既便捷又安全,生活中支付宝的出现使人们对现金的依赖逐渐弱化。随着智能手机的普及化和信息网络技术的应用,人们可以更容易地运用银行数字货币支付服务。中国手机普及率

为 94.5 部每百人，而只有 64% 的人拥有银行账户，银行可以积极开拓大量无法获得银行账户但通过互联网对接的客户。其中一个途径就是，通过数字货币建立数字钱包，在金融覆盖不足和经济欠发达地区实现更低成本、更安全的小额支付和资金转移，实现中间业务收入增加。

3. 区块链

工信部 2016 年《区块链发展白皮书》中将区块链定义为一种分布式数据存储、点对点传输、共识机制、加密算法等计算机技术在互联网时代的创新应用模式。区块链是一种融合多种现有技术的新型分布式计算和存储范式。它利用分布式共识算法生成和更新数据，并利用对等网络进行节点间的数据传输，结合密码学原理和时间戳等技术的分布式账本保证存储数据的不可篡改，利用自动化脚本代码或智能合约实现上层应用逻辑。如果说传统数据库实现数据的单方维护，那么区块链则实现多方维护相同数据，保证数据的安全性和业务的公平性。区块链通过程序和代码，将规则内嵌到计算机系统中，且没有处于核心的可以随意篡改数据的管理人员，它是一种提供信任的社会基础设施。

区块链技术可以显著改变银行风险管理方式。2014 年高盛报告的数据表明，全球银行为了满足监管部门的要求，不断增加对合规部门系统安全的资金投入和人力资本的投入以建立完善的信用机制和征信机制，相关投入共计 180 亿美元。同时，银行在建立反洗钱系统的过程中，还需不断对客户信息进行多次核实与调查，其无形花费无法准确估量。如果银行等传统金融机构将区块链技术应用到反洗钱等风险管理系统中，依靠其账本式分布和无法篡改的时间戳等特性可以最大限度降低银行的维护成本与人力成本。区块链的可追溯特征可以保证任何一笔交易资金与操作步骤永久地保留在区块链系统中，能够有效避免因监管不完善与法律漏洞所引发的洗钱活动。区块链技术的共享性可以保证系统中的每一个节点都可以由各节点随时查找与检查，一方面可以减少银行对信息的复查，提升银行运行效率，另一方面金融信息的共享可以帮助银行等金融机构快速、准确地寻找到合适的客户，减少银行的运营成本。此外，区块链的去中心化特征可以提升银行的金融安全，区块链系统中的每一个节点都无法完全掌握、修改系统中的信息，并且每个节点的关键数据都将以私钥的形式存储，不仅可以防止信息被篡改，也可以避免信息的泄露。根据高盛的研究估算，如果银行等金融机构将区块链技术应用到反洗钱等风险防范计划中，那么全球金融机构对于风险安全的投资将减少 20 亿～30 亿美元，其中减少的人工成本为 1.6 亿美元，金融交易与数据审查可节省 14 亿美元，而在系统的优化与保护方面可节省 5 亿美元。

区块链技术影响银行结算体系。当前，电子交易与结算收入已经成为传统银行的重要业务收入来源。在实际电子交易过程中，银行为了解决电子资金不能即时清算的问题，将第三方金融中介引入结算系统中，使交易的发生成为可能。虽然第三方金融中介为资金的结算提供了便利，但是大量的第三方金融机构介入银行的结算系统中，无形增加了银行的运营成本，降低了运营效率并伴有大量金融风险。而区块链结算系统与传统银行结算系统相比，在成本、安全与运营效率等方面优势明显，具体表现如下。

第一，用户在使用银行的结算系统时，必须持有银行卡或者输入账号与密码才可以进行资金操作，之后再由商家与银行进行验证从而完成交易。在这一过程中，客户所输入的信息有可能被不法分子获取从而造成不必要的损失。而区块链结算系统可以通过哈希加密值的方法免除客户的基本金融信息输入，区块链技术会将客户的基本信息映射为二进制代码，并

一同验证信息的准确性与真实性。

第二,区块链结算系统通过分布式账本的形式,使区块链系统中的每一个参与人实现信息共享,消除不必要的第三方中介,降低银行交易手续费。以美国 Stripe 为例,该公司的区块链支付系统中设定交易额小于 100 万美元的资金不收取手续费,而超过 100 万美元的部分将按照 1%的费率收取手续费,远低于当前银行的手续费率。

第三,区块链支付系统可以缩短交易时间,提升运营效率。传统银行借记卡与信用卡交易基本都是在 1 个工作日内完成,跨境支付转账更是需要 3～5 个工作日的时间。而区块链支付系统的 P2P(点对点)交易方式与去中心化的特征可以保证资金瞬时到达,增强资金的流动性。对于大额度的跨境支付交易,缩短交易时间更是可以防止因汇率发生改变而造成的金融损失。

4. 边缘计算

随着万物互联时代的快速到来和无线网络的普及,网络边缘的设备数量和产生的数据都快速增长。在这种情形下,以云计算模型为核心的集中式处理模式将无法高效处理边缘设备产生的数据。在万物互联的背景下,传统的云计算存在实时性不够、宽带不足、能耗较大、不利于数据安全和隐私等问题。为了解决这些问题,面向边缘设备所产生海量数据计算的边缘计算模型应运而生。边缘计算是一种将计算、存储、网络资源从云平台迁移到网络边缘的分布式信息服务架构,试图将移动通信网、互联网和物联网等业务进行深度融合,减少业务交付的端到端时延,提升用户体验。边缘计算中边缘的下行数据表示云服务,即将传统云计算中心的服务下移到边缘设备上执行,上行数据表示万物互联服务,而边缘计算的边缘是指从数据源到云计算中心路径之间的任意计算和网络资源。边缘计算模型和云计算模型并不是取代关系,而是相辅相成的关系,边缘计算需要云计算中心强大的计算能力和海量存储的支持,而云计算也需要边缘计算中边缘设备对海量数据及隐私数据的处理。

边缘计算模型具有几个明显的优点:首先,在网络边缘处理大量临时数据,不再全部上传云端,这极大地减轻了网络带宽和数据中心功耗的压力。其次,在靠近数据生产者处做数据处理,不需要通过网络请求云计算中心的响应,大大减少了系统延迟,增强了服务响应能力。最后,边缘计算不再上传用户隐私数据,而是将其存储在网络边缘设备上,减少了网络数据泄露的风险,保护了用户数据安全和隐私。

由于上述优势,边缘计算的发展前景良好,已经在社会生活的多个领域得到了应用,例如城市公共安全中实时数据处理便采用了边缘计算技术。随着智慧城市和平安城市的建设,大量传感器被安装到城市的各个角落,提升公共安全。例如,武汉的“雪亮工程”建设于 2019 年 6 月底,实现了全市公共安全视频监控总量达到 150 万个。得益于“雪亮工程”的建设,全市刑事有效警情同比下降 27.2%,并为群众查找走失老人小孩、追回遗失贵重物品等服务 1 万余次。随着共享经济的兴起,各种共享经济产品落地并得到发展,如滴滴、Uber 和共享单车。然而,这些产品同时也存在大量的公共安全事件。例如,顺风车司机对乘客进行骚扰,甚至发生刑事案件。因此,2018 年 9 月,受顺风车安全事件的影响,滴滴已经临时下线顺风车业务,并进行整改,首当其冲的是在司机端加入服务时间段的自动录音功能。然而,想要进一步提升安全性,最终还是得依赖于视频等技术,但这将导致大量的带宽需求。按照 Uber 2017 年的使用情况(45787 次每分钟),假设将每次驾乘的视频发送至云端(每次 20min),每天云端将新增 9.23PB 的视频数据。边缘计算作为近数据源计算,可以大幅度降

低数据带宽，将可以用来解决公共安全领域视频数据处理的问题。虽然当前城市中部署了大量的摄像头，但是大部分摄像头都不具备前置的计算功能，而需要将数据传输至数据中心进行处理，或者需要人工的方式来进行数据筛选。然而在边缘计算技术能够实现视频的实时处理，同时实现和周边摄像头的联动。

5. 物联网

现如今，全世界的计算机网络已经成为一张巨大的网，地球仿佛就像一个小村庄。在这个小村庄里发生了任何事，如果有人发表在网络上，那世界各地的人们都可以通过浏览去了解。计算机网络作为物联网存在的基础，物联网想要再向上发展必须依靠计算机网络技术。物联网的发展也变得越来越智能化和小型化，这也对信息网络技术的速度、覆盖率等提出了更高的要求。

物联网的发展实现了万物互联。物联网是指把各种信息传感设备（如射频识别器、全球定位系统、激光扫描器、宏伟感应器）等装置和互联网结合组成的网络。它的功能是让物品通过网络连接起来，让人们能够很容易地管理和识别它们。物联网通过计算机网络技术建立，其中 RFID 系统是物联网的基础。它结合了当前的网络技术、数据库和中间件技术等，由无数联网的阅读器和移动标签组合而成，构成了一个比 Internet 更庞大的物联网。通过物联网，系统可以随时随意地对物体进行识别、追踪、定位以及监控。其体系结构共分为以下三层。

（1）感知层。物联网的组成模块中，最基础的模块便是感知层，物联网技术的主要功能基本也是靠它来完成的，它主要是去感知网络区域的数据，然后通过数据或者资料的查询，再经过网关传输查找到用户需要的数据信息。网关的功能是连接全部网络，再把得到的信息进行处理。要让计算机完成所有的信息操作和处理操作，需要让计算机变得更加完善、能力更强，同样也需要应对更多的任务。随着物联网的加速发展，网络的信息和数据量变得十分庞大，这给各方面的工作都带来了很大的难度，如数据搜集、分析和处理等。物联网的感知层要获得信息感知必须进行大量的工作才可以完成。

（2）应用层。其作用是将处理完的信息传达给用户，通过对信息的分析，再选取用户需要的信息数据。物联网可以通过优化计算机网络配置等，去挖掘计算机网络的最大功能，尽可能地满足人们的要求。物联网应用层的功能为物联网的发展提供支撑，从另一角度来看，它关系着计算机网络技术的发展。物联网虽然是独立的个体，但是它对计算机的影响却很大。应用层是物联网技术的基础，也是物联网发展的最大推力，它把所有的资源集中在一起，通过网络技术来进行平台搭建，从而优化物联网的发展环境，平衡计算机网络和物联网之间的矛盾。

（3）传输层。计算机网络技术是物联网传输层的基础，宽带实现了通信网络和感知层之间的连接，有效地把传输层的作用发挥出来。在传统的通信网络中，物联网传输层整合了所有节点并进行统一管理，发挥了十分关键的作用。只是在现阶段，各个通信节点都是相互独立的个体，都具有隐私权。物联网技术除了传输数据信息外，还成了连接物与物之间的一座桥梁，因为这个原因，让物联网的工作程序变得十分繁杂，而且物联网和它的工作空间都必须具有独立性，所以这也是对物联网工作的和计算机网络技术的发展造成了一定的影响。

当下物联网在现代社会的应用已越来越广泛，其典型代表便是城市智能交通系统。智能交通系统是指将先进的传感器技术、信息技术、网络技术、自动控制技术、计算机处理技术

等应用于整个交通运输管理体系从而形成的一种信息化、智能化、社会化的交通运输综合管理和控制系统。智能交通系统使交通基础设施能发挥最大效能。智能交通是一个综合体系,它包含的子系统大体可分为以下几个方面。

(1) 车辆控制系统。它是指辅助驾驶员驾驶汽车或替代驾驶员自动驾驶汽车的系统。该系统通过安装在汽车前部和旁侧的雷达或红线探测仪,可以准确地判断车与障碍物之间的距离,遇紧急情况时,车载计算机能及时发出警报或自动刹车避让,并根据路况自己调节行车速度,人称“智能汽车”。

(2) 交通监控系统。该系统类似于机场的航空控制器,它在道路、车辆和驾驶员之间建立快速通信联系,哪里发生了交通事故、哪里交通拥挤、哪条路最为畅通,该系统会以最快的速度将这些信息提供给驾驶员和交通管理人员。

(3) 运营车辆高度管理系统。该系统通过汽车的车载计算机、高度管理中心计算机与全球定位系统卫星联网,实现驾驶员与调度管理中心之间的双向通信,来提供商业车辆、公共汽车和出租汽车的运营效率。该系统通信能力极强,可以对全国乃至更大范围内的车辆实时控制。

(4) 旅游信息系统。它是专为外出旅行人员及时提供各种交通信息系统。该系统提供信息的媒介是多种多样的,如计算机、电视、电话、路标、无线电、车内显示屏等。无论在办公室、大街上、家中、汽车上,只要采取其中任何一种方式,都能从信息系统中获得所需要的信息。有了该系统,外出旅游就可以眼观四路、耳听八方了。

6. 无线通信技术

智能通信设备的普及,使人们对无线通信技术的要求越来越高。移动无线通信技术是指在移动通信网络、无线传输设备等要素支持下所形成的一种技术。该技术应用过程中不仅可支持语音通信,也能满足文字及多媒体的传输要求。同时,在多样化标准的支持下,移动无线通信技术的应用范围正在扩大,可为数据信息的高效传递、通信质量提高等提供技术保障。

无线通信技术经历了多代的发展,以不断满足人们的高要求。如今,5G 时代的到来,极大地方便了人们的即时通信,也给许多行业的运营模式、人们的生产生活方式等带来了非常大的改变。与传统的通信模式相比,5G 移动通信在速度、效率、稳定程度等方面都有很大的优势,它的性能目标是高数据速率、减少延迟、节省能源、降低成本、提高系统容量和大规模设备连接。

5G 将成为构筑经济社会的重要基础设施,5G 正从移动互联向万物互联拓展,随着中国 5G 在 2019 年正式步入商用,各行各业已在积极培育应用产业;5G 需要攻克实体经济数字化转型的难关,迫切需要 5G 产业与各行业一起构建 5G 产业新生态,通过培育 5G 先行应用的重点行业,推动 5G 应用场景、解决方案、产品及商业模式的发展。

(1) 5G+直播。直播是互联网技术发展到一定阶段的产物,直播过程中的流畅度、用户观感和良好用户体验对于网络的时延、宽带和稳定性均有较高的要求。随着 4K、8K、虚拟现实/增强现实(VR/AR)等超高清、交互性直播方式的发展,网络和设备性能将面临更大的挑战。5G 的发展将给超高清视频直播带来强大的网络支撑,这依赖于 5G 具有传输速率更快,能够解决高清视频直播中的卡顿问题;流量密度更大,可以保障用户在体育场、大型购物场所、交通枢纽等人员密集区域开展视频直播需求;可靠性和抗干扰能力强,能保证视

频直播的稳定性；时延率极低，可以提供几乎实地的视频直播等特点。5G 网络逐步向超高清视频直播渗透，目前已形成了 5G＋直播的示范性应用，且 5G 将打破现有视频直播面临的网络性束缚，推进视频直播的发展。

虎牙直播依托 5G 网络实现了 5G＋4K 高清户外直播。虎牙直播与中国电信合作，首次在直播行业进行了 5G 商用探索，结合 5G 与边缘计算开展高清视频直播业务尝试，顺利完成了 5G＋4K 高清户外直播实验，成为中国率先实现 5G 网络直播的平台。

音乐盛典咕咪汇依托 5G 实现了 4K 直播全过程应用。中国移动和华为在音乐盛典咪咕汇上首次成功实现了 5G 网络切片在全球大型直播活动中的应用，实现了 5G＋4K 直播从拍摄、编/转码到传输等全过程。主要应用场景包括两个方面：一是在红毯、主舞台等地方利用多台摄像机拍摄超高清直播信号，通过 5G 网络切片实时上传到咪咕视讯云数据中心进行制作和分支；二是利用 5G 网络切片接收后期制作的信号，在现场 4K 显示屏上对颁奖礼进行全程直播。

武汉大学“樱花节”实现 5G 超高清视频直播。武汉大学在“樱花节”联合湖北移动和中兴通讯，在校园中架设 15 个拍摄机位并配备 360°高清全景摄像头，现场采集的超高清画面通过 5G 终端、5G 基站和 5G 核心网实时上传到武汉大学视频服务器，并同步传送至新华网、人民日报、抖音、咪咕等平台进行直播，从而实现 5G＋4K 超高清视频直播。

两会期间山东省利用 5G＋VR 实现全景直播。山东联通携手省内多家主流媒体利用 5G 和 VR 全景技术对“两会”进行了全景 VR 直播。通过在会场内部安装的 VR 摄像头对视频信息进行了专业化采集，利用 5G 网络实时将信号传送回山东。观众通过微信平台便可以直击两会现场，享受低延迟、高质量带来的极为细致的视觉盛宴。

(2) 5G＋云游戏。5G＋云游戏是指游戏主体在云端服务器运行，通过 5G 网络传输游戏画面、音频和控制信息，实现流畅清晰的用户游戏体验。5G 网络依靠其网络宽带大、时延低等特性，能满足 5G＋4K 超高清云游戏的需求。5G＋云游戏的高速率特性使得游戏下载时间大幅度减小，无须等待，即点即玩；其云端处理能力降低了用户手机终端性能要求；它还实现了云端存储，在终端的游戏大小可降至 10MB，无须占用大量的手机存储空间。

浙江移动依托 VR 实现 5G＋云游戏。浙江移动推出的 5G 云 VR 方案是全国首次基于 5G 试验网下开通的云 VR 业务，其业务包含四大精品业务类型：“足不出户的实景直播”“1MAX 超宽巨幕的影视体验”“亲临现场的 8K-VR 现场直播”“身临其境的 VR 云游戏”。

咪咕互动娱乐推出 5G 云游戏产品推广计划。2018 中国移动全球合作伙伴大会上，咪咕互娱向行业公布了“5G 快游戏”的产品推广计划，并对“5G 快游戏”的技术优势和商业价值进行了深入解读。5G 正在加速部署，游戏已经走向云端，超高清互动、沉浸式体验的全场景游戏时代终于到来了。咪咕互娱的“5G 快游戏”是基于云游戏技术的下一代游戏平台，能够为用户带来随时、随地、任意设备的全场景沉浸式游戏体验，让用户随时随地地畅玩各类游戏。

(3) 5G＋360°全屏。5G＋360°全屏是将 5G 传输和 VR/AR 技术有机结合的应用。360°VR/AR 是借助近眼现实、感知交互、渲染处理、网络传输和内容制作等新一代信息技术构建的超越端、管、云的新业态，可让用户有亲临现场般的体验。360°VR/AR 有着非常广阔的应用价值和未来市场，但 360°VR/AR 技术对整个通信过程的网络性能有较高的要求，并直接关系用户实际体验，如该技术需要低时延来避免用户体验中出现的眩晕感、需要

高宽带来支撑高清镜头采集的高清内容传输。5G 技术有着百兆级宽带、毫秒级时延,其正式商用为 360°VR/AR 的需求提供了有力保障,也促进了 5G+360°全屏的应用融合。

中国移动依托 5G+360°全屏实现了对水乡景色的 VR 直播。第五届世界互联网大会上,中国移动推出了业界首个基于 5G 网络传输的 8K VR 实时直播。中国移动在直播方案中采用深圳看到科技研发的 Obsidian 专业 VR 相机以及 8K 3D 全景直播软件 Kandao Live 8K,将实际风景以 8K 分辨率实时展现在 110 英寸的大屏幕上。

2019 年央视春晚实现了 5G+360°全屏的 VR 直播。2019 年中央电视台春节联欢晚会上,中国联通、华为与中央电视台合作,在中央广播电视总台布放 5G 室内数字化设备,推出央视超高清视频 VR 直播,为观众带来了不一样的感受体验。

2021 年 10 月,脸书(Facebook)宣布将公司名称更改为元(META),引发了"元宇宙"这一概念的热度以及国内外诸多企业的关注。"元宇宙"这个概念目前还没有较为完备的定义,其最早起源于 1992 年著名的美国科幻小说家尼奥·斯蒂文森(Neal Stephenson)撰写的《雪崩》(*Snow Crash*)。书中描述了一个平行于现实世界的网络世界——元宇宙(Metaverse)。所有现实世界中的人在元宇宙中都有一个网络分身(Avatar),可以随时随地切换身份,自由穿梭于物理世界和数字世界,在虚拟空间和时间节点所构成的元宇宙中学习、工作、交友、购物、旅游等。依据书中的描述,理想的元宇宙需要通过 VR 技术构造一个逼真的虚拟世界。此外还需要对现实世界中的万事万物乃至人类采集信息,将一切结构化、非结构化数据统统封装为特定格式的元数据上传至网络,这一步需要通过结合 VR 和物联网技术来实现。最后人们能够在元宇宙中实现自由地学习、工作和交易等行为,还用到了人工智能和基于区块链的数字货币技术。因此无论最终如何定义,元宇宙本质上都是人工智能、区块链、VR 等前沿技术的综合性应用。

3.3.2 信息网络重构社会新形态

随着科技水平的提高,信息网络技术也在不断地创新和更加符合人们的各种需求,人们的生活越来越依赖于信息网络技术,反过来信息网络技术的发展也不断推动着社会的进步。信息网络技术对社会的影响主要体现在以下 5 个方面。

1. 社会关系网络化演变

近年来,社会关系的研究已成为社会网络科学领域最热门的课题之一。信息网络的发展突破了固有的时空限制,信息媒介的突变能力使得这种连接既可以将具有不同地域、空间的个体联系在一起,也可以将非个人的、非正式的社会组织连接起来。大数据时代,社会关系伴随着信息网络的不断延伸也在发生着颠覆式的改变,数字世界和物理世界的跨越超越了传统基于血缘、信仰、爱好、民族、国籍的限制,每个人、每个物都成为社会关系演化过程中的一个中间体,从简单网络结构到复杂网络结构、从边界清晰到界限模糊、从高度组织化到常态不确定性、从高度集中到分布式、从封闭到全球化,社会关系的网络化演化不可逆转。在农业社会,等级制度决定了社会关系结构的线性化;工业化时代,面临日益增长的复杂性和不确定性,社会关系的单一结构也被多维线性关系所取代,但是,不可否认,这种多维线性结构依然是线性的。而在信息网络推动下的大数据时代,网络结构跨越了时空限制,突破原有多维线性结构的局限,多元化合作共存成为社会关系演化的主旋律。多元化的结构不仅

意味着连接主体的多元化，也包括连接方式的多元化、连接关系的多元化。

信息网络时代下的每个个体并不是孤立存在的，社会群体的表现形式也十分丰富，可以是家庭、同学、朋友、工作单位、相同爱好的群体，甚至是志愿服务组织，每个人都或多或少与各类群体有着千丝万缕的联系。根据美国社会学家马克·格兰诺维特 1973 年发表的《弱关系的力量》一文中所指出的：传统社会由于传播渠道和范围的有限，社会关系表现为“强关联”，而在信息网络背景之下，数字网络取代物理网络成为信息分享和传递的主要渠道，“弱关系”虽然不像“强关联”的社会关系那样坚固，但是其低成本和高效率的信息传播比“强关联”更容易跨越社会结构层级的界限去获取信息，进出创造出更多的“弱关系”。处于网络中的人们既不是毫无关联的一盘散沙，也不是休戚相关的小圈子，而是在开放和互动中保持一种空间、时间、功能上的有序演化。

互联网信息分发模式可以大致分为四个时代：启示初代网民上网可以做什么的门户网站时代；基于关键词主动查找感兴趣的信息的搜索引擎时代；用户参与门槛更低、参与程度更高的社交媒体时代；已经基于用户画像精准投放产品的推荐算法时代。每个时代，信息分发的速度、精度、量级都要远超前一个时代。

在传统的门户网站时代，网站拥有最大的话语权和信息资源，虽然网络论坛（BBS）的出现实现了公众有限的话语权，但并未形成气候。而到了移动互联网时代，用户的角色开始发生变化，可以更加自由地在网络中浏览及分享信息。信息传播的成本降低且效率剧增。自媒体、社会性媒体、社交网络等新概念层出不穷，彻底改变了媒体和信息传播的方式，任何人都可以利用社会化媒体来实时传播身边的第一手信息，这直接颠覆了过去由主流媒体一统天下的格局。传统的传播者-接收者泾渭分明的界限被彻底打破，“话语平权”成为一种可能性，传统大众媒体在传统社会中所拥有的风光在社会化网络时代已经不复存在。

2. 重构新经济增长模式

在工业社会中，经济增长主要依靠加大物质资源投入的方式来实现，这种粗放型的增长方式存在诸多弊端，如会引起不可再生资源枯竭、加剧环境污染等。而在信息网络技术催生的新经济环境下，增长的实现主要依靠信息咨询、知识、智慧和科技创新，摆脱了高投入、高消耗、高污染的经济发展方式，不断催生出新的产业和服务，开拓出一条创新创意、资源节约、环境友好、效率惊人的新路。目前，互联网在不同国家对 GDP 的直接贡献范围为 0.8%～6.3%，在过去五年里互联网对经济增长的贡献超过 20%，而主要的推动力就是网络经济。例如，小米手机采用互联网线上营销模式，而且非常注重用户体验，它不仅卖手机，还提供增值服务，后续可以不断通过应用服务获取利润。小米手机以异于传统手机公司的商业模式，借助互联网创新其商务模式，促进自身的经济增长。

1）共享经济

共享经济最早是 1978 年由美国得克萨斯州立大学社会学教授马科斯·费尔逊（Marcus Felson）和伊利诺伊大学社会学教授琼·斯潘思（Joel Spaeth）在其发表的论文中提出，与传统经济方式不同之处在于共享经济是一种依托现代信息技术，对机构和个人分散、闲置资源进行再配置，用以满足多样化社会需求，提升资源使用效率为目标的经济模式。共享经济的组成要素包括交易对象、交易平台和交易主体。交易平台作为连接交易主体供需方的纽带，通过整合闲散资源，满足交易主体各方个性化需求。

共享经济并不是新生事物，传统方式下人们之间的互借行为也是一种形式的共享，但此

种方式的共享受限于空间和关系的限制：一是共享双方只能在其所触达的空间进行交易；二是双方基于点对点的相互信任关系才能达成。而在大数据时代，信息网络技术的发展为共享经济的发展开拓了思路和拓展了范围，交易主体不再受制于空间限制，通过去中介化的交易平台实现交易对象的再中介化。其中，去中介化是指共享经济的出现，打破了交易服务提供方对组织和机构的依附，他们可以直接向最终服务需求方提供服务或产品；再中介化是指交易服务提供方虽然脱离组织和机构的依附，但为了更广泛地接触需求方，他们接入互联网的共享经济平台。

信息网络的发展为共享经济提供了更大机遇，共享出行、共享住宿、共享金融、共享物流等共享平台的出现，通过撮合交易，实现资源和物品所有权的短暂转移；反过来，供给方和需求方的相互促进又使得共享经济不断壮大，使得分散的交易行业具备了更大规模的可能性。

2）新实体经济

实体经济是国家社会生产力的直接体现，新兴技术的应用，互联网＋的无限拓展使得传统的实体经济企业面临巨大的压力和挑战。新实体经济最初是由阿里巴巴集团马云先生提出的一个经济概念，如今已被经济界、学术界所认可。新实体经济是传统行业与新型信息网络技术结合产生的新经济形式，传统的旧经济建立在制造业的基础上，以标准化、规模化、模式化、讲求效率和层次化为其特点；而新经济则是建立在信息技术基础上，追求的是差异化、个性化、网络化和速度化。新实体经济的本质是信息化和全球化，过去传统的实体经济严重依赖政府调控来减少市场行为的盲目性，而新实体经济通过信息化和网络化将越来越多的传统企业纳入新经济的生态圈中。

在过去的十几年，互联网在国内的蓬勃发展，加速了信息网络向传统产业的渗透，互联网＋已深入人心，它带来的不仅是高效率的信息处理能力，更包括跨行业的产业协同和全球范围内的市场融合。交通、医疗、消费、安全、教育等各个行业正在快速地发生着改变，新的消费需求，新的商业模式将层出不穷。

3. 突破传统的产业区位选择机制

传统区位理论在进行区位选择时，模型的核心变量是运输成本、市场需求等，新经济地理学把空间集聚的规模收益等因素加入到区位选择研究中，从地理距离上升到空间布局层次。广播、电视、互联网和其他信息媒介的出现，使得人与人之间的时空距离骤然缩短，整个世界紧缩成一个“村落”。交通工具便捷化推动社会网络远程化，通信技术和互联网技术快速发展推动信息获取和交流更加高效化，全球化的国际贸易扩展了商业和社会的互联性。各类社交工具(如微信、QQ、微博等)的应用使得交往和沟通变得触手可及，一部手机、一台计算机就能实现跨地区、面对面的交流，信息网络技术与现代交通工具的便捷化使得城市和地区之间的物流频率得到显著提升，城市不再是人和工业的聚集，而是社会网络的中心，跨国间的互动交流更加频繁，信息全球化促进生产全球化和贸易全球化进入更高的层次，国际间的协作分工更为精细化，产品可以来自世界各地，消费可以多种多样，物理空间的间隔已被信息网络的优势所弥补，社会化协作由封闭走向更加开放。

在网络经济发展模式下，效率和创新成为决定企业存在的关键性因素。一方面，信息技术对于经济运行中驱动利润高低的因素从“空间运输成本”变为“时间成本”，通常表现为企业对于市场的快速反应以及产品的快速配送；同时网络空间正在侵蚀现实空间，造成地域

无差别的现象。另一方面，网络改变着人们的消费方式和认知，使得个性化需求的满足成为更重要的目标，这就要求顺应潮流的知识、技术、设计创新，昭示着新的“区位”选择趋向。例如诺基亚与黑莓衰落的事实已向世界展示了全球市场里的新“游戏规则”。可以说，在新的区位选择论中，核心变量应当是区位的经济运行效率和有效创新能力。效率主要受区域信息网络设施、产业基础、规章制度等的影响；而创新则是能够顺应市场潮流，满足不断产生的新的需求的创新，这项能力的大小由区域内集中的人才素质所决定，包括人员的市场洞察力、思维灵活度、研发能力等。当然，这些新的因素既突破了传统的地理、空间范围的束缚，同时又表现出与之高度的相关性。可以说是在理论的基础上，引领新的方向。

4. 改变经济活动的不公平性

根据马克思主义社会形态理论，生产力的发展仅作用于生产关系，是生产关系向更高维度转变的关键因素。原始狩猎社会生产力水平低下，生产资料公有、集体劳动决定了社会关系的高度集权化。奴隶社会以奴隶主占有奴隶的人身自由为主要特征的生产关系和农业社会以土地所有制为核心的生产关系决定社会关系的以奴隶主和地主为核心的集权控制体系。封建社会在生产力方面是很落后的，农民的生产力水平很低，工业生产规模很小，而商业由于与农业生产的矛盾被统治阶级所抑制。因此，注定了封建社会是以农业为主，手工业和商业为辅的社会形态。而到了工业社会和信息社会，生产力技术水平的大幅提升，原本的附着于奴隶主和地主的集权制度受到生产力的冲击。社会内部个体之间、个体与群体之间、群体与群体之间的关系和属性也在发生着新的改变。

在以往的经济活动模式中，各种生产要素、基础条件的差异常常会导致不同市场主体（厂商、个人、地区）参与经济活动的不平等性。而在信息网络高速发展的市场环境下，市场的进入门槛大大降低，即使那些偏远的、条件差的地区，也会由于信息网络的延伸而改变命运，打通交流的渠道，将会有厂商主动为该地区或主体提供原本缺失的要素，将其纳入大市场中来。例如农村电商，在农村建设网络后，可以利用网络电商平台，扩大当地独具特色、品质优异的农特产品的销售渠道，并通过农村电商为当地农村的基础设施建设提供更多的资金支持。拼多多农村电商扶贫项目为“社交扶贫＋拼多多”模式，依托社交关系推进电商，促进同类兴趣的细分顾客聚集，帮助农产品更加容易突破销售瓶颈。这一模式具体过程为：通过预售制提前聚起海量订单，再把大单快速分解成大量小单，直接与众多农户对接，优先包销贫困户家中农货，实现在田间地头“边采摘、边销售”。

5. 颠覆传统的信息网络安全体系

当人工智能可以更加便捷地做到数据资源的保护，在入侵检测技术的人工智能进入入侵检测应用之后，能够将数据与安全领域专家的经验相结合，建立起人工智能推理机制并同时对网络特征编码进行预处理，更新数据库，找出符合编码特征的信息，以此判断入侵原因，进行入侵危害检测以及类似入侵防护。在入侵检测方面应用基础上搭建人工智能神经网络，对数据进行智能化分析，进行数据的整合分类，最后根据整合结果建立具有针对性的拦截数据库以及信息过滤操作，以此能够显著提升网络数据安全性。

例如，垃圾邮件是人们工作、生活当中使用电子邮箱时会经常收到的，垃圾邮件的出现对于人们正常使用邮箱有一定的阻碍性，人工智能在对邮箱进行监测的同时能够筛选出符合条件的垃圾邮件并将这些邮件拦截不让其进入收件箱内。这样做可以帮助用户纯粹地使

用邮箱而不受到垃圾邮件的干扰。

信息识别过程中可以使用人工智能技术,模糊处理存在的不确定信息。

3.3.3 信息网络赋能产业数字化

随着互联网的不断发展,数字世界与物理世界的结合越来越紧密,信息网络技术的重心已经从消费互联网向产业互联网、价值互联网转化,区块链是这个过程中的可信数据基础设施和金融建设基础设施,有力支持各行业的数字化转型发展。各地政府和众多企业开始在区块链领域寻找新的业务突破口,产业区块链已经成为区块链行业发展的主战场。同时,在国家大力推动"新基建"[①]的大环境下,数字化资产将成为企业的重要资产,数据搜集、上链、算法分析和模型运用、数据交易流转,以指导经营和促进产业发展,数字化转型是大势所趋。

1. 消费互联网

消费互联网是以个人为用户,以日常生活为应用场景的应用形式,满足消费者在互联网中的消费需求而生的互联网类型。消费互联网以消费者为服务中心,针对个人用户提升消费过程的体验,在人们的阅读、出行、娱乐、生活等诸多方面有很大的改善,让生活变得更方便、更快捷。消费互联网的本质是个人虚拟化,增强个人生活消费体验。

1) 发展历程

消费互联网的发展先由 PC 端向移动端转移,由一线城市用户向农村用户普及。消费互联网的市场竞争格局和产业格局趋向于成熟和稳定。以电商为例,在 B2C 市场,2017 年,天猫与京东两大巨头占据了 80%以上的份额,双寡头格局明显。近几年,阿里巴巴、京东的成交总额增速开始放缓。

人口红利结束之后,消费互联网已经呈现饱和状态,各线上行业渗透率已经接近天花板。随着消费互联网的成熟,互联网逐渐由消费向产业发展。在未来相当长的时间里,基于现有技术,进行产业互联网和消费互联网的结合,用产业互联网提升产业的效率,来改善消费互联网的用户体验。

随着用户、数据和支付的统一,未来的创业趋势将是围绕一群有共同属性的用户,不断挖掘他们的需求,进入不同的行业,提供全方位的产品和服务。

2) 主要特征

消费互联网以提供个性娱乐为主要方式,在短时间内迅速吸引眼球,但由于其服务范围的局限性,且未触动消费者本质生活,也易导致其迅速淹没于互联网发展的浪潮中。消费互联网依托于强大的信息与数据处理能力,以及多样化的移动终端的发展,在电子商务社交网络、搜索引擎等行业出现规模化发展态势,并形成各自的生态圈,奠定了稳定的行业发展格局。

3) 商业模式

消费互联网以"眼球经济"为主,通过高质量的内容和有效信息的提供来获得流量,从而

---

① 新型基础设施建设(简称新基建)在 2020 年 3 月首次被提出,主要包括 5G 基站建设、特高压、城际高速铁路和城市轨道交通、新能源汽车充电桩、大数据中心、人工智能、工业互联网七大领域,涉及诸多产业链,是以新发展理念为引领,以技术创新为驱动,以信息网络为基础,面向高质量发展需要,提供数字转型、智能升级、融合创新等服务的基础设施体系。

通过流量变现的形式吸引投资商，最终形成完整的产业链条。腾讯、百度、今日头条等都是典型的消费互联网公司，它们面向个人用户提供产品和体验，并借助聚集起来的巨大流量，直接（游戏付费等）或者间接（广告等）地获得收入。

在消费互联网时代，个人消费者是主要服务对象，提供个性娱乐是主要服务方式，流量变现是主要商业模式，即通过高质量的内容和有效信息的提供来获得流量，从而通过流量变现的形式吸引投资，最终形成完整的产业链条。

一大批消费互联网垂直领域企业合并，如滴滴和快的、58 和赶集、美团和大众点评、携程和去哪儿等，标志着行业发展已到一定阶段，开始从深度整合走向成熟。

4）消费互联网的成功因素

消费互联网创新的低成本进入带来了很多创新参与者。尽力而为是互联网发展初期的宗旨。它不需要先期的市场研究、产品设计、市场推广。几个人搭起草台，依靠自己的 IT 能力，编写自己认为好的应用，在互联网这个现成的平台上发布，直接测试市场的需求。它允许早期的低质量，在不断迭代下，通过在网上和客户的互动，抓取客户的需求，从而不断提升业务和应用的质量和体验。互联网应用的发展是通过赛马机制来完成的。好的应用自然得到保留，得不到市场认可的应用自然被淘汰。大数量级的创新奠定了消费互联网快速发展的基础。

消费互联网发展的高速及巨大的规模效应特征成为金融投资极好的标的物。投资买的是未来，泡沫是金融市场的特征。消费互联网指数级的增长及全球市场的预期，给金融市场以很好的打造投资标的机会。在初期亚马逊采用收费模式不久就被 Yahoo 的免费模式击败。而其背后支撑其发展的经济力量正是华尔街。华尔街以其独有的对互联网发展趋势的判断，迅速培育、推升了互联网泡沫，吸引大量投资加入。蜂拥而入的资金趟平了互联网发展初期消费者进入所面临的价值判断的高壁垒，快速形成规模市场。

双边市场成就互联网的经济落地，接住了互联网泡沫。互联网发展初期的免费最终还是需要市场买单。这就是 2000 年互联网泡沫破灭的原因。但幸运的是广告市场形成了消费互联网的双边市场模式，很好地支撑起消费互联网的发展。在 2000 年互联网泡沫破灭之后，Google 创立了 Adwords 广告印钞机模式。通过对广告内容的自动获取、投放以及收费和进一步的匹配优化，以极低的成本实现了企业的快速盈利，进而带动了互联网发展的黄金时代。直到现在，互联网盈利的三大支柱，广告、游戏、电商，广告仍然是其重要的收入来源，且电商中的一大部分也是来源于广告。目前很热的网红带货模式，也不过是广告的变种。但这同时也带来了消费互联网发展中要面临的问题。到目前为止，除了广告这一双边市场的成功之外，并没有新的双边市场出现。这就预示着未来的发展大概率要回归传统模式，即消费者付费模式。摆在我们面前的一个典型案例就是共享单车，在经历了 40 亿元的疯狂风投之后，共享单车终于走回了向用户收取骑行费的传统模式。

另外，还有一个不为人关注但却不可或缺的要素就是消费互联网的目标人群——年轻人，其以极低的学习成本进入互联网，消除了新产品采用中的重要障碍。

但对于年轻人甚至于未成年的孩子使用互联网都毫无障碍。这一批人成为了互联网创新应用忠实的追随者，为互联网的发展壮大起到了不可忽视的作用。而产业互联网就需要培育其员工，改变他们的习惯，这面临的将是文化和模式的变革。

2. 产业互联网

产业互联网是以企事业单位为主要用户、以生产经营活动为关键内容、以提升效率和优化配置为核心主题的互联网应用和创新,是数字经济深化发展的高级阶段。它促进数字世界与物理世界的打通,能够使产业的组织方式、商业模式、运作流程等各个方面发生显著的改变,并由此提高生产效率和经济效益。它是基于互联网技术和生态,对各个垂直产业的产业链和内部的价值链进行重塑和改造,从而形成的互联网生态和形态。产业互联网是一种新的经济形态,利用信息技术与互联网平台,充分发挥互联网在生产要素配置中的优化和集成作用,实现互联网与传统产业深度融合,将互联网的创新应用成果深化于国家经济、科技、军事、民生等各项经济社会领域中,最终提升国家的生产力。

随着产业互联网的实践推进,出现了各类由区域政府或者产业骨干企业打造的产业互联网平台。由于各产业平台发起背景和资源能力优势的不同,因此其发展路径也有所差异。

1) 行业龙头企业的裂变式增长

大型行业龙头企业发起推动的产业互联网平台,将过去在产业积累的客户、人才、技术等方面的综合资源优势和核心能力通过平台开放化,打造产业级生产性服务业共享平台,为产业链上下游企业进行赋能,以大企业带动产业链中小企业共同发展,实现产业链整体转型提升,同时自身也在传统业务之外打造出一家基于互联网的新模式公司,实现裂变式增长。

2) 区域特色产业集群的转型升级

以区域政府、行业协会或产业骨干企业多方共同发起打造产业互联网平台,带动区域产业集群的整体转型升级,将成为推进县域经济创新发展的重要手段。这类产业互联网实践具有鲜明的县域产业集群特色,通过产业链的打通实现产业的融合。县域特色产业集群往往由当地政府支持行业协会中的骨干企业以及当地国有投资控股企业、金融和投资机构等联合发起,具有熟悉产业生态、掌握产业关键资源要素、易获得投资等天然优势,也更容易得到政策倾斜、孵化期资源支持等,但同时需要避免发生架构不稳定、落地执行效果差等问题。要保证这类平台的健康发展,必须设立合理的公司市场化运作股权架构和治理体系,同时考虑对于核心管理团队的激励机制。

3) 专业商贸市场的数字化转型

专业商贸市场具有天然的平台优势,以及丰富的产业资源,通过数字化转型,将线下客户资源优势与线上平台一体化融合打通,可以为产业链上的从业者提供从交易、支付,到物流、供应链金融等领域的供应链专业服务,通过线上交易数据的累积,为交易双方提供信用保证体系,促进交易双方的强粘性服务,提升复购率和交易效率,大大降低交易成本,推动整个产业生态的提升。

4) 商贸/物流商到供应链集成服务商转型

在传统产业链中提供贸易、物流等服务的企业,基于过去比较好的品牌影响力、线下资源等优势积累,正在进一步向产业供应链的集成服务商转型。商贸/物流商到供应链集成服务商转型,其关键成功要素是从全产业链的视角对于产业场景需求和痛点的挖掘,在前期需做好产业互联网的顶层设计规划。

5) 行业资讯平台/SaaS 解决方案商的产业互联网升级

在早期互联网的发展过程中,涌现出一批行业资讯平台,往往名称为“XX 网”,为行业圈子提供行情资讯、价格指数等,积累了大量的行业用户信息和流量。由于缺乏服务深度和

粘性，往往难以为继，因此纷纷转型产业互联网，从提供撮合交易到产业链的集成服务。还有另一类行业 SaaS 解决方案提供商，基于行业大数据的优势积累，通过大数据的分析应用，进一步往产业供应链服务延伸。

产业互联网是服务于生产的互联网，在产业互联网时代主要以生产者为主体，实现所有行业、企业、生态链关系和企业迭代周期的互联网化。也就是说，产业互联网就如同人类的生产力从蒸汽时代迈入电气时代的发展，是生产力脱胎换骨的改变。在不久的将来我们所能罗列的制造业、教育、农业、医疗行业、交通、运输以及市政管理甚至公务员行业都会被互联网化，随之而来的将是企业的生产方式、组织运营方式、产业边界和商业模式的巨变。

消费互联网与产业互联网的区别主要有两个：一个是用户主体不同，消费互联网针对的是个人用户，产业互联网针对各行各业的生产者；另一个是兴起动因不同，消费互联网的目的在于满足人们的某些生活体验和消费需求，产业互联网的目的在于通过生产、资源整合实现快速发展。

可以说，产业互联网是消费互联网的进一步发展和深化，而它也将再一次改变人类社会的生活方式和发展历程。

3. 价值互联网

价值互联网是一个新兴的概念，在信息网络成熟之后，特别是区块链的出现，为价值互联网带来了新的发展空间，触发了一个新的发展阶段。价值互联网是以区块链技术为核心基础，依托移动设备与数字货币实现价值交换与价值存储，且其作为新型金融技术，每年的投资额正以倍数形式增长。

1）初步发展阶段

广义上讲，价值互联网的雏形可以追溯到 20 世纪 90 年代，美国安全第一网络银行(SFNB)在 1996 年开始网上金融服务，中国在 1998 年也有了第一笔网络支付。其后，很多金融机构借助互联网技术来拓展支付业务，并出现了第三方支付、大数据金融、网络金融门户等模式，以互联网金融为代表的价值互联网相关产业不断发展，价值互联网特征逐渐显现。尤其是 2010 年以来，随着互联网金融呈现爆发式增长，价值互联互通的范围不断扩大，程度逐渐提高，价值互联网的功能有了初步发展。

2）全网发展阶段

区块链的出现，为价值互联网带来了新的发展空间，触发了一个新的发展阶段。可以说，在区块链出现之前，价值互联网处于一个非常初级的发展阶段，基本上是以一些中介化机构为中心的碎片化发展模式。而区块链具有去中心化、透明可信、自组织等特征，使得其应用更容易扩散为全球范围内的无地域界限的应用，为价值互联网注入了新的内涵。随着区块链应用的逐渐发展，将推进形成规模化的、真正意义上的价值互联网。

互联网技术解决了信息不需要通过第三方便可以实现数据信息在全球的高效流通的问题。而现在互联网架构中再建立一套价值传递的机制成为新一轮互联网发展的动力。区块链技术的诞生就是为了实现价值互联网的建立。区块链是一种在对等网络环境下，通过透明和可信规则，构建不可伪造、不可篡改和可追溯的块链式数据结构，实现和管理事务处理的模式。区块链是分布式数据存储、传输、加密等计算机技术在互联网时代组合创新的应用模式，具有分布式对等、数据块链式、不可伪造和防篡改、透明可信、高可靠性等特征，被视作

大型机、PC、互联网之后计算模式上的又一次颠覆创新,正在推动信息互联网向价值互联网转变,有望改变财税金融、贸易流通、生产制造、社会管理等人类社会活动形态。

区块链赖以生存的基础就是高信用度高安全性以及点对点的现金支付模式。分布于全球的各个节点共同来存储交易数据,共识机制保证了区块链网络的安全性。在区块链系统中,不需要第三方机构的参与,直接实现点对点传输的交易。

在金融领域里面,区块链可以用于支付、借贷、保险、众筹等。在政治事务方面,区块链可以用于选举、身份验证、公共信息的查询、项目招标等。在物流方面,区块链可以用于供应链的数据的保护、信息追踪、信息监测等使得物流行业透明化。在公益慈善方面,区块链可以实现点对点的捐助,善款信息便得透明可以追溯。在农业领域,可以使得从农田到餐桌整个流程上面的信息可查,来保证食物安全问题。

这几年,区块链领域得到了很大的发展,以区块链为名的公司正在不断崛起。一类公司是平台型的,主要提供区块链技术的操作系统,然后给企业提供服务,方便它们在系统上面去开发自己想要的应用。另一类公司就是应用类型的公司。另外,互联网的巨头们也在不断越来越深地摄入区块链领域的研究,如阿里巴巴、腾讯、百度、迅雷等。另外,各大银行也开启了区块链项目的研发。

不过目前来说,很多公司都是初创公司,这就意味着,区块链技术的成功还有很长的一段路要走。价值互联网是互联网技术由信息互联网发展的必然方向。信息传播的快速发展使人们的生活发生了翻天覆地的变化。而基于区块链技术的价值互联网相信必将会更深程度地影响人们生活的方方面面。

3.4 信息网络技术

通过本章内容的学习,已经了解到利用信息网络技术可以从 Web 中获取大量的信息。那么如何采集这些信息进而为大数据分析提供所需要的数据呢? 最简单、直接的方法就是用 Python 的网络爬虫(Crawler)技术来解决。在本节中,将介绍网络爬虫的相关知识,引导读者使用 Python 语言构建网络爬虫并获取网络中的数据。


视频讲解

3.4.1 网络爬虫基础知识

1. 网络爬虫的概念与分类

人们通过浏览器来浏览网页,而网络爬虫是通过模仿浏览器来访问网页,它可根据某种规则自动获取所需要的网络信息。使用 Python 可以很方便地编写出爬虫程序,进行互联网数据的自动获取。爬虫又可分为通用爬虫和聚焦爬虫。其中,通用爬虫就是人们每天使用的搜索引擎“抓取系统”的重要组成部分。其主要目的是将互联网上的网页下载到本地,形成一个对互联网已发布内容的镜像备份。通用爬虫会尽可能地把互联网上的所有的网页下载下来,放到本地服务器中形成备份,再对这些网页做进一步处理(如提取正文、去掉广告),最后提供一个用户检索接口。而聚焦爬虫是根据指定的需求抓取网络上指定的数据。例如:获取电影的名称和演员,而不是获取整张页面中所有的数据。聚焦爬虫会按照设定的规则,自动地抓取网页中的信息,并能沿着网页的相关链接在网络中采集资源,是一个功

能很强的网页自动抓取程序。

目前网络爬虫已被广泛应用于搜集 Web 网页、文档、图片、音频、视频等资源。网络爬虫主要分成 4 个步骤：①发送请求；②获取响应内容；③解析内容；④保存数据，如图 3-8 所示。

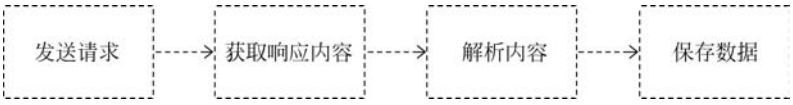


图 3-8　网络爬虫主要分成的 4 个步骤

2. HTML 简介

要从互联网中提取有用的数据，还需要了解用于创建的网页的标准标记语言：超文本标记语言（Hyper Text Markup Language，HTML）。HTML 也被称为网页源代码，它是一种通过标签来描述网页的语言，由标签和文本内容与属性构成。

HTML 标签是由大括号包围的关键词组成的，例如< html >。HTML 标签通常是成对出现的，如< b >和。标签对中的第一个标签是开始标签，第二个标签是结束标签。开始标签和结束标签也被称为开放标签和闭合标签。而开始标签和结束标签之间的文本被称为标签内容，如< p >这是标签内容</p >。

人们使用的网页浏览器（如 Chrome、Internet Explorer、搜狗、Safari 等）便是用于读取 HTML 文件，并将其内容显示出来的软件。如果需要用户查看 HTML 的源代码，以 Chrome 浏览器为例，可以通过在浏览器窗口右击，在弹出的快捷菜单中选择“查看网页源代码”命令，如图 3-9 所示。最终，查看网页源代码的结果通过 HTML 标签展示，如图 3-10 所示。

图 3-9　选择“查看网页源代码”命令

事实上，HTML 标签可转换为一棵 HTML 树，如图 3-11 所示。该树也被称为 DOM（Document Object Model）树，它是一种层次模型。DOM 树将网页中的各个元素都看作一

图 3-10 查看网页源代码的结果通过 HTML 标签展示

个个对象,对象处于某个层次中,从而使网页中的元素也可以被计算机语言获取或者编辑。DOM 是以层次结构组织的节点或信息片断的集合,DOM 树把 HTML 文档呈现为带有元素、属性和文本的树结构。这个层次结构允许开发人员在树中导航以寻找特定信息。

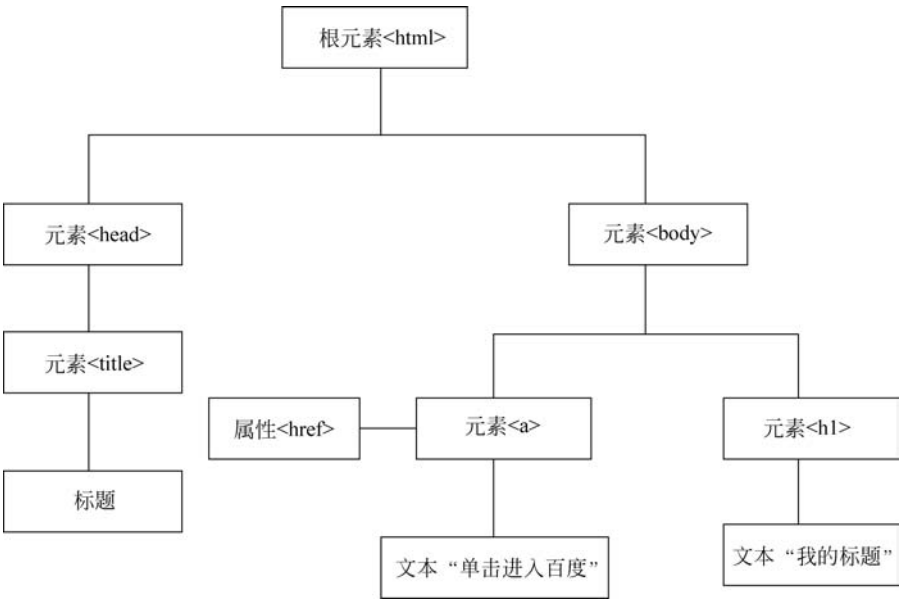


图 3-11 HTML 标签可转换为一棵 HTML 树

3. requests 库的安装和使用

利用 Python 语言获取互联网上的 HTML 源代码首先需要安装第三方库——requests 库,requests 库的作用就是请求网站获取网页数据的,Python 的第三方库可以通过 pip 命令来安装。

在 cmd.exe 窗口中输入 pip 命令,如图 3-12 所示,如果返回 pip 命令的使用方法,说明 pip 命令可以正常使用。

如果出现提示“‘pip’不是内部或外部命令,也不是可运行的程序”错误信息,则说明 Python 环境变量没有设置好,需要修复或者重新安装。

输入 pip install requests 命令完成 requests 库的安装,如图 3-13 所示。注意,安装的过程中需要计算机处于联网状态,这样才能从网站下获取第三方库。

图 3-12　在 cmd.exe 窗口中输入 pip 命令

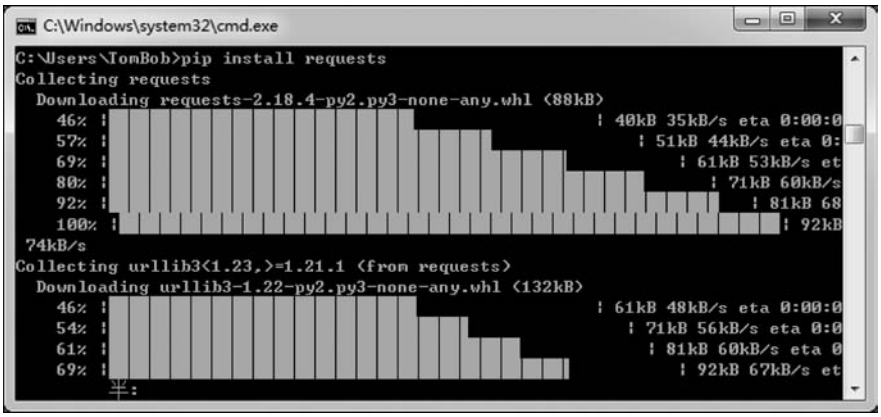


图 3-13　输入 pip install requests 命令完成 requests 库的安装

在安装成功 requests 库之后，通过 requests 库输入网址获取网页内容。使用 requests 库获取网页内容最基本的方法是 get()请求。例如获取访问百度主页的代码如下：

```
>>> import requests
>>> url = "http://www.baidu.com"
>>> res = requests.get(url)
```

通过 requests.get()方法返回的是一个 response 对象，上面将其保存在 res 变量中，可以通过 res 变量来查看 response 对象的属性，其代码如下：

```
>>> res.status_code
200
>>> res.encoding
'ISO-8859-1'
>>> res.encoding = 'utf-8'
>>> res.text
```

```
<!DOCTYPE html>
<!--STATUS OK--><html><head><meta http-equiv=content-type content=text/html;
charset=utf-8><meta http-equiv=X-UA-Compatible content=IE=Edge><meta content=
always name=referrer><link rel=stylesheet type=text/css href=http://s1.bdstatic.
com/r/www/cache/bdorz/baidu.min.css><title>百度一下,你就知道</title></head><body
link=#0000cc><div id=wrapper><div id=head><div class=head_wrapper><div class=
s_form><div class=s_form_wrapper><div id=lg><img hidefocus=true src=//www.
baidu.com/img/bd_logo1.png width=270 height=129></div><form id=form name=f action=//
www.baidu.com/s class=fm><input type=hidden name=bdorz_come value=1><input type=
hidden name=ie value=utf-8><input type=hidden name=f value=8><input type=
hidden name=rsv_bp value=1><input type=hidden name=rsv_idx value=1><input type=
hidden name=tn value=baidu><span class="bg s_ipt_wr"><input id=kw name=wd class=
s_ipt value maxlength=255 autocomplete=off autofocus></span><span class="bg s_btn_
wr"><input type=submit id=su value=百度一下 class="bg s_btn"></span></form></div>
</div><div id=u1><a href=http://news.baidu.com name=tj_trnews class=mnav>新闻
</a><a href=http://www.hao123.com name=tj_trhao123 class=mnav>hao123</a><a href=
http://map.baidu.com name=tj_trmap class=mnav>地图</a><a href=http://v.baidu.com
name=tj_trvideo class=mnav>视频</a><a href=http://tieba.baidu.com name=tj_trtieba
class=mnav>贴吧</a><noscript><a href=http://www.baidu.com/bdorz/login.gif?
login&amp;tpl=mn&amp;u=http%3A%2F%2Fwww.baidu.com%2f%3fbdorz_come%3d1 name=
tj_login class=lb>登录</a></noscript><script>document.write('<a href="http://
www.baidu.com/bdorz/login.gif?login&tpl=mn&u='+encodeURIComponent(window.location.
href+(window.location.search===""?"?":"&")+"bdorz_come=1")+'" name="tj_
login" class="lb">登录</a>');</script><a href=//www.baidu.com/more/ name=tj_
briicon class=bri style="display:block;">更多产品</a></div></div></div><div
id=ftCon><div id=ftConw><p id=lh><a href=http://home.baidu.com>关于百度</a>
<a href=http://ir.baidu.com>About Baidu</a></p><p id=cp>&copy;2017 
Baidu <a href=http://www.baidu.com/duty/>使用百度前必读</a> <a href=
http://jianyi.baidu.com/ class=cp-feedback>意见反馈</a> 京ICP证030173号
 <img src=//www.baidu.com/img/gs.gif></p></div></div></div></body>
</html>
```

在上述代码中,res.status_code 表示 response 对象的状态代码,200 表示连接成功。res.encoding 表示 response 对象内容的编码方式。可以通过代码 res.encoding = 'utf-8',将原来的 ISO-8859-1 编码方式修改为 UTF-8 编码方式,这样可以正常地显示中文字符。最终,通过 res.text 显示通过爬虫获取到的网页内容。

通过 requests.get()方法返回的 response 对象中,包含一些常用的属性来表征请求响应后的结果。response 对象的常用属性及说明如表 3-5 所示。

表 3-5 response 对象的常用属性及说明

属 性	说 明
res.status_code	HTTP 请求的返回状态代码
res.text	URL 对应的网页内容
res.encoding	HTTP 响应内容的编码方式
res.content	HTTP 响应内容的二进制形式
res.json()	requests 中内置的 JSON 解码器
res.raise_for_status()	失败请求(非 200 响应)抛出异常

4. 网络爬虫需要遵守的协议

网络爬虫可从网络服务器抓取各种信息，其中可能存在涉及个人隐私或商业机密的内容，如果将不合适的内容提供给爬虫使用者，可能会给服务器管理者带来不必要的困扰与纠纷，因此需要 Robots 协议(Robots Exclusion Standard)来对网络爬虫进行规范。Robots 协议是网络爬虫排除标准，其作用为网站告知网络爬虫哪些页面可以抓取，哪些网页不可以抓取。该协议的内容存放于网站根目录下的 robots.txt 文件中。例如，京东的 robots 文件网址为 https://www.jd.com/ robots.txt，其内容如下：

```
User-agent: *
Disallow: /?*
Disallow: /pop/*.html
Disallow: /pinpai/*.html?*
User-agent: EtaoSpider
Disallow: /
User-agent: HuihuiSpider
Disallow: /
User-agent: GwdangSpider
Disallow: /
User-agent: WochachaSpider
Disallow: /
```

User-agent: * 中的 * 代表的所有的搜索引擎种类，* 是一个通配符，代表所有；Disallow: /?* 禁止访问所有以问号开头的链接；Disallow:/pop/*.html 代表禁止访问 pop 目录下以 html 结尾的网址；Disallow:/pinpai/*.html?* 代表禁止访问 pinpai 目录下 html 后面接? 号的网址。

当用 requests.get()方法访问网页时，get()方法后面还可以设置一个 headers 参数。当爬虫访问服务器时，有些服务器检查访问 headers 的 User-Agent 域，如果发现不是浏览器访问，服务器会拒绝访问。这时，可以手工加入 User-Agent 域，加入某种浏览器的请求头信息，其代码如下：

```
res = requests.get(url, headers = {'User-Agent':'Mozilla/5.0'})
```

上述代码将模拟 Mozilla/5.0 浏览器访问相应的 URL 网页。

3.4.2　Python 网络爬虫实战


视频讲解

1. 基于正则表达式的数据获取

正则表达式(Regular Expression)是用于处理字符串的强大工具，通常被用来从文本中抽取符合某种规则的内容。Python 中使用正则表达式，只需要在程序前面加入 import re 即可。正则表达式拥有自己独特的语法，它们主要用于模式的匹配。其常见的语法如下所述。

(1) . 表示匹配任意字符，换行符\n 除外。

(2) * 表示匹配前一个字符 0 次或无限次。

(3)?表示匹配前一个字符0次或1次。

(4).* 表示贪心算法。

(5).*?表示非贪心算法。

(6)()表示括号内的数据作为结果返回。

在基于正则表达式的Python网络爬虫程序中,最常用的方法就是re.findall(pattern,string),该方法表示从string字符串中按照pattern这种模式去匹配内容。需要注意的是re.findall()方法的返回值为列表,之后还需要遍历这个列表来获取所需的内容。通过这种方法可以获取
标签中的内容,即所有学院名称。其代码如下:

```
import re
s = '''<html>
            <li>信息与安全工程学院</li>
            <li>新闻与传播学院</li>
            <li>法学院</li>
     </html>'''
results = re.findall('<li>(.*?)</li>',s)
for each in results:
     print(each)
```

其中,s为所需要的处理的字符串,因为其中有换行,所以用三引号将其引起来,re.findall(pattern,string)方法表示从string字符串中按照pattern这种模式去查找内容,方法的第一个参数代表匹配的模式,在上例中
(.*?)
表示按照左边是
、右边是
的模式从s中抽取字符串。在上例中满足左边是
、右边是
的恰好为所有学院的名称。

使用正则表达式的抽取方法主要就是观察需要抽取内容的左右两边,将需要抽取的中间内容用(.*?)代替即可,因为上例中的三个学院的名称都处于左边是
、右边是
的标签之间,所以用模式
(.*?)
就可以抽取出它们的内容。

2. 基于XPath的数据获取

XPath是XML Path Language的缩写,它是一种小型的查询语言,它既可以在HTML中查找信息,也可以通过元素和属性进行导航。第三方XPath库也需要先安装再使用,其安装方法为pip install lxml。XPath能够根据HTML的语法建立解析树,进而高效解析其中的内容。XPath使用路径表达式来选取HTML文档中的节点或者节点集。这些路径表达式和常规的计算机文件系统中看到的表达式非常相似。XPath的简单调用方法代码如下:

```
from lxml import etree
selector = etree.HTML(网页源代码)
selector.xpath(路径表达式)
```

上面的源代码为网页的HTML源代码,路径表达式为XPath语法规则,XPath主要有6种标签的使用方法。

(1)//(双斜杠),定位根节点,会对全文进行扫描,在文档中选取所有符合条件的内容,以列表的形式返回。

(2) /(单斜杠),寻找当前标签路径的下一层路径标签或者对当前路标签内容进行操作。

(3) /text(),获取当前路径下的文本内容。

(4) /@xxxx,提取当前路径下标签的属性值。

(5) .(点),用来选取当前节点。

(6) ..(双点),选取当前节点的父节点。

下面通过具体实例讲解通过 XPath 获取网页中的数据,其代码如下:

```
from lxml import etree
html = '''< html >
       < head >
          < title > Python 网络爬虫</title >
       </head >
       < body >
           < div class = 'university'>
              < a href = "http://www.zuel.edu.cn/">中南财经政法大学</a >
           </div >
         < div class = 'school'>
              < li >< a href = "http://xagx.zuel.edu.cn/">工程学院</a ></li >
              < li >< a href = "http://tsxy.zuel.edu.cn/">数学学院</a ></li >
              < li >< a href = "http://law.zuel.edu.cn/">法学院</a ></li >
         </div >
       </body >
</html > '''
```

【例 3-1】 通过 XPath 获取标题数据“Python 网络爬虫”。

实现代码如下:

```
page = etree.HTML(html)
tags = page. xpath ("//title/text()"))
for tag in tags:
     print(tag)            # 打印标题 Python 网络爬虫
```

在上面的代码中,通过//title 搜索根目录下的< title >标签,通过//title/text()获取< title >标签中的内容。上面代码的运行结果为:

```
Python 网络爬虫
```

【例 3-2】 通过 XPath 获取所有< a >标签中的内容和超链接网址。

实现代码如下:

```
page = etree.HTML(html)
links = page.xpath("//a/@href")
for link in links:
     print(link)           # 获取 href 的超链接
```

在上面的代码中,通过//a 搜索根目录下的< a >标签,通过//a/@href 获取所有的超链

接信息。上面代码的运行结果为：

```
http://www.zuel.edu.cn/
http://xagx.zuel.edu.cn/
http://tsxy.zuel.edu.cn/
http://law.zuel.edu.cn/
```

【例 3-3】 通过 XPath 获取< a >标签中的学院内容，而不包括学校，可以先采用//div[@class='school']获取学院所在的< div >标签，再获取< a >标签中的内容。

实现代码如下：

```
page = etree.HTML(html)
schools = page.xpath("//div[@class = 'school']/li/a/text()")
        #通过//div[@class = 'school'],搜索根目录下 class 属性值为 school 的< div >标签
for school in schools:
    print(school)
```

在上面的代码中，通过//div[@class='school']搜索根目录下 class 属性值为 school 的< div >标签，然后通过/li/a/text()搜索该< div >标签下< li >标签下面的< a >标签内容，其运行结果为：

```
工程学院
数学学院
法学院
```

使用 XPath 的一个优势在于在多数浏览器中都有审查元素的功能，如 Chrome 浏览器右键快捷菜单中的“审查元素”命令，如图 3-14 所示。在右边选择所需的 HTML 代码，然后右击，在弹出的快捷菜单中选择 Copy XPath 命令复制对应的 XPath，如图 3-15 所示。通过 Copy XPath 命令可以直接将该元素的 XPath 代码复制到粘贴板中。

图 3-14 Chrome 浏览器右键快捷菜单中的“审查元素”命令

3. 使用 Python 操作 Excel 表格

Python 可以采用 Python 自带的 csv 库将数据写入 Excel 表格中。操作 Excel 表格时需要将 Excel 数据转换为一个二维列表，如图 3-16 所示。

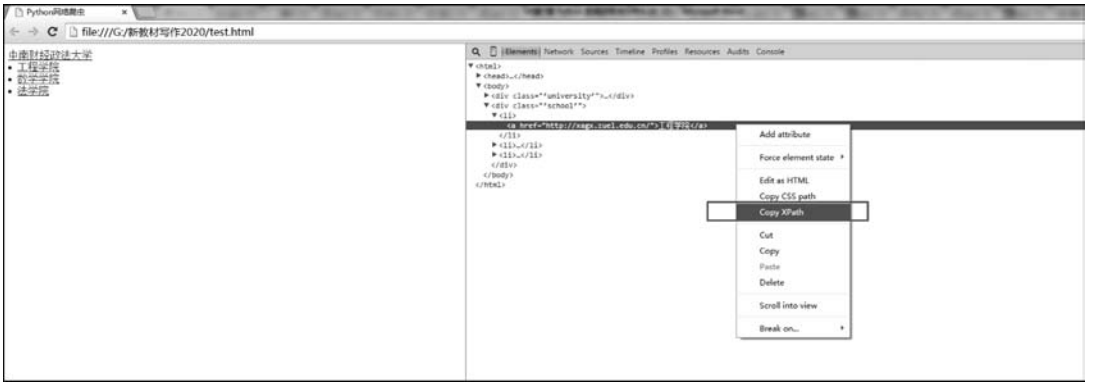


图 3-15　选择 Copy XPath 命令复制对应的 XPath

A	B	C
D	E	F

⇨

```
rows=[
['A','B','C'],
['D','E','F'],
]
```

图 3-16　操作 Excel 表格时需要将 Excel 数据转换为一个二维列表

在上例中，rows=[['A','B','C'],['D','E','F']]实际上对应的就是 Excel 中第一行是 A,B,C,第二行是 D,E,F 的表格，在写入数据时，需要用 open()方法建立一个 csv 文件并打开它，其代码如下：

```
file = open('test.csv','w',newline = '')
```

open()函数的第一个参数 test.csv 是需要打开的文件名字，第二个参数 w 代表以写入的方式打开文件，第三个参数 newline=''表示在写入时行与行不需要空行。Python 写入 Excel 的代码如下：

```
import csv                                  #导入 csv 库
rows = [['A','B','C'],[ 'D','E','F']]       #定义二维列表
file = open('test.csv','w',newline = '')    #以写入的方式打开 test.csv 文件
f_csv = csv.writer(file)                    #准备写入
f_csv.writerows(rows)                       #写入数据
file.close()                                #关闭文件
```

4. 基于正则表达式获取中南映像数据案例

下面介绍通过 Python 获取中南映像数据(网址为 http://www.zuel.edu.cn/2019n/list.htm)，如图 3-17 所示。

首先，需要提取标题，通过对 HTML 源文件的观察，获得标题的模式与代码，如图 3-18 所示。

上例中的粗体部分“25 万字，一份来自中南大的知识产权强国战略专家意见”是需要从网页源代码中抽取的内容，只需要把粗体部分替换为(.*?)就变成所需要的模式，最后在

图 3-17　通过 Python 获取中南映像数据

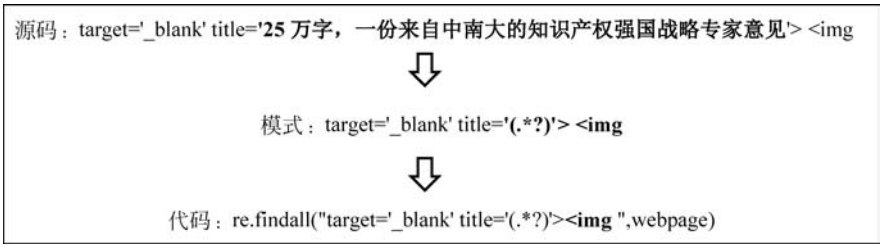


图 3-18　获得标题的模式与代码

Python 代码中使用 re.findall("target='_blank' title='(.*?)><img ",webpage),将会把 webpage(HTML 源代码)中所有满足模式条件的字符串以列表的形式返回。

接着,要分析提取供稿单位的 HTML 源代码,获得供稿单位的模式与代码,如图 3-19 所示。

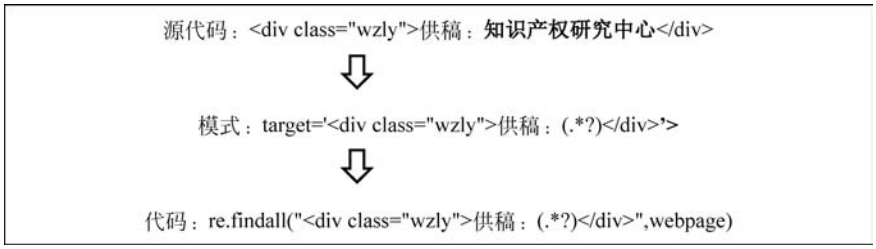


图 3-19　获得供稿单位的模式与代码

上述粗体字部分“知识产权研究中心”是需要从网页源代码中抽取的内容,只需要把粗体部分替换为(.*?)就可以变成所需要的模式。在 Python 中使用 re.findall(" 供稿：(.*?) ",webpage),将会把 webpage(HTML 源代码)中所有满足模式条件的字符串以列表的形式返回。

通过 Python 获取中南映像数据的完整代码如下所示：

```
import re                                                    #导入 re 库
import requests                                              #导入 requests 库
url = 'http://www.zuel.edu.cn/2019n/list.htm'                #设置 2019 年中南映像的网址
response = requests.get(url)                                 #访问 2019 年中南映像的网址
response.encoding = 'utf-8'                                  #设置编码格式为 utf-8
webpage = response.text                                      #获取 2019 年中南映像网页源代码
titles = re.findall("target = '_blank' title = '(.*?)'><img",webpage)  #获取所有的标题信息
feeds = re.findall('<div class = "wzly">供稿：(.*?)</div>',webpage)  #获取所有的供稿信息
print("所有标题：")
for title in titles:                                         #依次打印所有的标题
    print(title)
print("所有供稿单位：")
for feed in feeds:                                           #依次打印所有的供稿单位
    print(feed)
```

最终程序的返回结果为所有的文章标题以及所有的供稿单位，返回结果内容如图 3-20 所示。

```
RESTART: C:/Users/Administrator/AppData/Local/Programs/Python/Python37/code/7基于正则表达
式获取中南映像数据案例.py
所有标题:
25万字，一份来自中南大的知识产权强国战略专家意见'
【身边的榜样】指数经济创新团队党小组：党旗下的育才故事'
希贤好故事(23) 韩翼：把热情洒满新疆这片土地'
【身边的榜样群体】教学督导：做中南大教学质量的“提灯人”'
石智雷：“学术研究的终点在于回馈学生”'
“研中学”“做中学” 培养卓越会计人才'
“教书育人奖”获得者熊波：用诗句讲解大学数学的女老师'
“教书育人奖”获得者孙贤林：为学生义务补课19年'
“教书育人奖”获得者王淑红：坚守初心的“投资顾问”'
“教书育人奖”获得者曹亮：教学能手 学生良师'
“教书育人奖”获得者黎江虹：先做益友 后为良师'
“教书育人奖”获得者刘惠好：把知识贡献给学生和社会'
“教书育人奖”获得者赵兴罗：将选修课变成“必修课”的财政史老师'
“教书育人奖”获得者卢现祥：新制度经济学的“传播者”'
所有供稿单位:
知识产权研究中心
统计与数学学院
校报学通社
教学督导与评估中心
公共管理学院
会计学院
统计与数学学院
会计学院
工商管理学院
工商管理学院
法学院
金融学院
```

图 3-20　返回结果内容

5. 基于 XPath 获取中国银行股票数据案例

获取股票数据对于金融分析非常有用，下面的案例将分析中国银行 2019 年第一季度的股票交易数据(网址为 http://quotes.money.163.com/trade/lsjysj_601988.html?year=2019&season=1)，如图 3-21 所示。

中国银行(601988)历史交易数据

2019 一季度 查询 下载数据

日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(手)	成交金额(万元)	振幅(%)	换手率(%)
2019-03-29	3.71	3.79	3.71	3.77	0.06	1.62	1,791,175	67,257	2.16	0.09
2019-03-28	3.72	3.73	3.69	3.71	-0.02	-0.54	1,087,031	40,306	1.07	0.05
2019-03-27	3.73	3.75	3.72	3.73	0.01	0.27	1,058,660	39,510	0.81	0.05
2019-03-26	3.75	3.75	3.71	3.72	-0.01	-0.27	1,250,035	46,574	1.07	0.06
2019-03-25	3.78	3.79	3.73	3.73	-0.07	-1.84	1,880,807	70,767	1.58	0.09
2019-03-22	3.80	3.82	3.78	3.80	0.01	0.26	1,419,561	53,923	1.06	0.07
2019-03-21	3.80	3.82	3.79	3.79	0.00	0.00	1,713,724	65,228	0.79	0.08
2019-03-20	3.81	3.82	3.77	3.79	-0.02	-0.52	1,648,942	62,576	1.31	0.08
2019-03-19	3.84	3.85	3.80	3.81	-0.03	-0.78	1,601,599	61,221	1.30	0.08
2019-03-18	3.79	3.84	3.77	3.84	0.05	1.32	2,130,221	81,025	1.85	0.10
2019-03-15	3.79	3.82	3.78	3.79	0.00	0.00	1,518,035	57,629	1.06	0.07
2019-03-14	3.80	3.83	3.77	3.79	-0.01	-0.26	1,289,656	48,960	1.58	0.06
2019-03-13	3.83	3.84	3.78	3.80	-0.03	-0.78	1,591,609	60,540	1.57	0.08
2019-03-12	3.79	3.87	3.79	3.83	0.04	1.06	2,473,383	94,854	2.11	0.12
2019-03-11	3.74	3.81	3.74	3.79	0.04	1.07	1,981,058	74,496	1.87	0.09

图 3-21　分析中国银行 2019 年第一季度的股票交易数据

通过分析发现该网页的数据有两种不同的格式,其中表头的 HTML 格式如图 3-22 所示。

```
<tr class="dbrow">
    <th>日期</th>
    <th>开盘价</th>
    <th>最高价</th>
<th>最低价</th>
    <th>收盘价</th>
    <th>涨跌额</th>
    <th>涨跌幅(%)</th>
    <th>成交量(手)</th>
    <th>成交金额(万元)</th>
    <th>振幅(%)</th>
    <th>换手率(%)</th>
</tr>
```

XPath:

//tr[@class="dbrow"]/th/text()

图 3-22　表头的 HTML 格式

对于以上 HTML 源代码采用 XPath 的路径表达式//tr[@class="dbrow"]/th/text()来解析里面的内容,其中 XPath 的前面部分//tr[@class="dbrow"]表示在整个网页中搜索 class="dbrow"的<tr>标签,XPath 的后面部分/th/text()表示在该<tr>下面寻找所有的<th>标签中的内容。

另外,表中数据的 HTML 格式如图 3-23 所示。

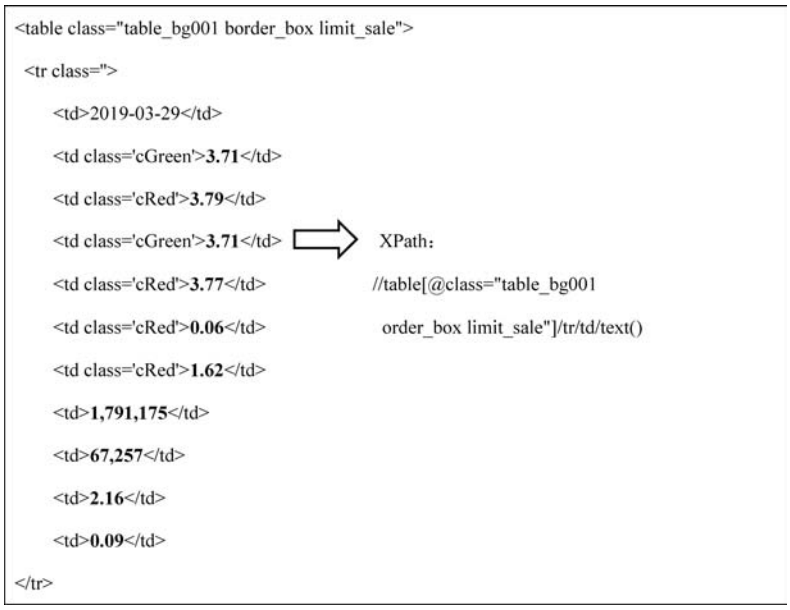


图 3-23　表中数据的 HTML 格式

其中，XPath 的前面部分//table[@class="table_bg001 border_box limit_sale"]表示在整个网页中搜索 class="table_bg001 border_box limit_sale"的< table >标签，XPath 的后面部分/tr/td/text()表示在该 table 里找到所有的< tr >下面的< td >数据，也就是所有的股票数据。

使用 Python 获取中国银行 2019 年第一季度的股票交易数据的完整代码如下：

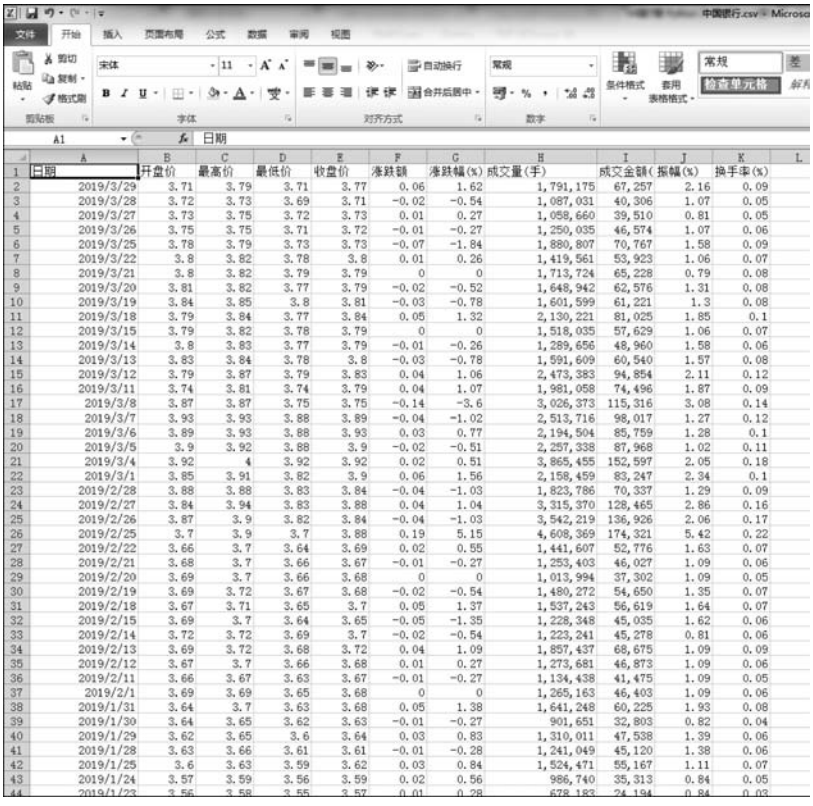
```
import requests
import re
from lxml import etree
import csv
url = "http://quotes.money.163.com/trade/lsjysj_601988.html?year=2019&season=1"
                                        #设置待抓取的股票网址
colnum = 11                             #股票一共有 11 列属性
res = requests.get(url)                 #访问股票网址
res.encoding = 'utf-8'                  #设置为 utf-8 编码
content = res.text                      #获取股票网页源代码
page = etree.HTML(content)
row = []                                #存储每一行数据
rows = []                               #按行存储所有数据
heads = page.xpath('//tr[@class="dbrow"]/th/text()')
                                        #设置股票表头的 XPath
rows.append(heads)                      #把表头加入到第一行
tds = page.xpath('//table[@class="table_bg001 border_box limit_sale"]/tr/td/text()')
                                        #股票内容的 XPath
i = 1                                   #i 用于计数，每 11 个为一组加入列表
for td in tds:
    if(i%colnum!=0):                    #如果不是 11 的倍数，就直接加入 row 中
        row.append(td)
    else:                               #每 11 个为一组存放到 rows 中
        row.append(td)
```

```
            rows.append(row)
            row = []
        i = i + 1
f = open('中国银行.csv','w',newline = '')      #将数据写入 Excel 文件中
f_csv = csv.writer(f)
f_csv.writerows(rows)
f.close()
print(len(rows),"行数据写入完毕")
```

由于股票中的数据都是在< td >中,需要将其按照每 11 个为一组加入列表,于是采用下面的循环结构并结合 if 语句,代码如下:

```
i = 1                              #i 用于计数
for td in tds:
    if(i % colnum!= 0):            #如果不是 11 的倍数,就直接加入 row 中
        row.append(td)             #把股票数据加入列表 row 中
    else:                          #每 11 个为一组存放到 rows 中
        row.append(td)             #把股票数据加入列表 row 中
        rows.append(row)           #把一行数据加入 rows 中
        row = []                   #清空 row 列表,用于下一次的数据加载
    i = i + 1                      #计数器 i 增加 1
```

最终,程序将获取后的股票数据写入 Excel 文件中,如图 3-24 所示。



	A	B	C	D	E	F	G	H	I	J	K
1	日期	开盘价	最高价	最低价	收盘价	涨跌额	涨跌幅(%)	成交量(手)	成交金额(	振幅(%)	换手率(%)
2	2019/3/29	3.71	3.79	3.71	3.77	0.06	1.62	1,791,175	67,257	2.16	0.09
3	2019/3/28	3.72	3.73	3.69	3.71	-0.02	-0.54	1,087,031	40,306	1.07	0.05
4	2019/3/27	3.73	3.75	3.72	3.73	0.01	0.27	1,058,660	39,510	0.81	0.05
5	2019/3/26	3.75	3.75	3.71	3.72	-0.01	-0.27	1,250,035	46,574	1.07	0.06
6	2019/3/25	3.78	3.79	3.73	3.73	-0.07	-1.84	1,880,807	70,767	1.58	0.09
7	2019/3/22	3.8	3.82	3.78	3.8	0.01	0.26	1,419,561	53,923	1.06	0.07
8	2019/3/21	3.8	3.82	3.79	3.79	0	0	1,713,724	65,228	0.79	0.08
9	2019/3/20	3.81	3.82	3.77	3.79	-0.02	-0.52	1,648,942	62,576	1.31	0.08
10	2019/3/19	3.84	3.85	3.8	3.81	-0.03	-0.78	1,601,599	61,221	1.3	0.08
11	2019/3/18	3.79	3.84	3.77	3.84	0.05	1.32	2,130,221	81,025	1.85	0.1
12	2019/3/15	3.79	3.82	3.78	3.79	0	0	1,518,035	57,629	1.06	0.07
13	2019/3/14	3.8	3.83	3.77	3.79	-0.01	-0.26	1,289,656	48,960	1.58	0.06
14	2019/3/13	3.83	3.84	3.78	3.8	-0.03	-0.78	1,591,609	60,540	1.57	0.08
15	2019/3/12	3.79	3.87	3.79	3.83	0.04	1.06	2,473,383	94,854	2.11	0.12
16	2019/3/11	3.74	3.81	3.74	3.79	0.04	1.07	1,981,058	74,496	1.87	0.09
17	2019/3/8	3.87	3.87	3.75	3.75	-0.14	-3.6	3,026,373	115,316	3.08	0.14
18	2019/3/7	3.93	3.93	3.88	3.89	-0.04	-1.02	2,513,716	98,017	1.27	0.12
19	2019/3/6	3.89	3.93	3.88	3.93	0.03	0.77	2,194,504	85,759	1.28	0.1
20	2019/3/5	3.9	3.92	3.88	3.9	-0.02	-0.51	2,257,338	87,968	1.02	0.11
21	2019/3/4	3.92	4	3.92	3.92	0.02	0.51	3,865,455	152,597	2.05	0.18
22	2019/3/1	3.85	3.91	3.82	3.9	0.06	1.56	2,158,459	83,247	2.34	0.1
23	2019/2/28	3.88	3.88	3.83	3.84	-0.04	-1.03	1,823,786	70,337	1.29	0.09
24	2019/2/27	3.84	3.94	3.83	3.88	0.04	1.04	3,315,370	128,465	2.86	0.16
25	2019/2/26	3.87	3.9	3.82	3.84	-0.04	-1.03	3,542,219	136,926	2.06	0.17
26	2019/2/25	3.7	3.9	3.7	3.88	0.19	5.15	4,608,369	174,321	5.42	0.22
27	2019/2/22	3.66	3.7	3.64	3.69	0.02	0.55	1,441,607	52,776	1.63	0.07
28	2019/2/21	3.68	3.7	3.66	3.67	-0.01	-0.27	1,253,403	46,027	1.09	0.06
29	2019/2/20	3.69	3.7	3.66	3.68	0	0	1,013,994	37,302	1.09	0.05
30	2019/2/19	3.69	3.72	3.67	3.68	-0.02	-0.54	1,480,272	54,650	1.35	0.07
31	2019/2/18	3.67	3.71	3.65	3.7	0.05	1.37	1,537,243	56,619	1.64	0.07
32	2019/2/15	3.69	3.7	3.64	3.65	-0.05	-1.35	1,228,348	45,035	1.62	0.06
33	2019/2/14	3.72	3.72	3.69	3.7	-0.02	-0.54	1,223,241	45,278	0.81	0.06
34	2019/2/13	3.69	3.72	3.68	3.72	0.04	1.09	1,857,437	68,675	1.09	0.09
35	2019/2/12	3.67	3.7	3.66	3.68	0.01	0.27	1,273,681	46,873	1.09	0.06
36	2019/2/11	3.66	3.67	3.63	3.67	-0.01	-0.27	1,134,438	41,475	1.09	0.05
37	2019/2/1	3.69	3.69	3.65	3.68	0	0	1,265,163	46,403	1.09	0.06
38	2019/1/31	3.64	3.7	3.63	3.68	0.05	1.38	1,641,248	60,225	1.93	0.08
39	2019/1/30	3.64	3.65	3.62	3.63	-0.01	-0.27	901,651	32,803	0.82	0.04
40	2019/1/29	3.62	3.65	3.6	3.64	0.03	0.83	1,310,011	47,538	1.39	0.06
41	2019/1/28	3.63	3.66	3.61	3.61	-0.01	-0.28	1,241,049	45,120	1.38	0.06
42	2019/1/25	3.6	3.63	3.59	3.62	0.03	0.84	1,524,471	55,167	1.11	0.07
43	2019/1/24	3.57	3.59	3.56	3.59	0.02	0.56	986,740	35,313	0.84	0.05
44	2019/1/23	3.56	3.58	3.55	3.57	0.01	0.28	678,183	24,194	0.84	0.03

图 3-24 程序将获取后的股票数据写入 Excel 文件中

本章小结

网络科技的发展伴随着人类社会的每一次跃迁，以物联网、5G、区块链、人工智能为代表的新兴科技为网络技术提供新的发展动能。万物互联改变的不仅仅是连接方式，以智能化为形态的网络科技跨越了人与物、物与物的边界，逐渐演变成人类-技术共存体，它既拥有无与伦比的力量，也具有与时俱进的特点。本章以信息网络技术对主要对象，详细介绍了其起源发展、技术体系和运行机制等相关内容，并对大数据时代下的信息网络技术发展新动态进行了梳理，最后介绍了基于信息网络技术与 Python 程序设计语言的大数据获取方法。在本章的学习过程中，希望读者能够对信息网络技术有更深入、全面的了解，并结合自己所学专业领域，尝试利用信息网络提供的丰富数据资源完成数据获取操作，进而为后续大数据分析打下基础，解决与专业领域相关的问题。

通过网络爬虫工具获取分析所需的数据之后，接下来该如何处理？如何展示这些信息？在第 1 章中已经了解，大数据分析中涉及的数据类型繁多，常见的有结构化数据和非结构化数据。根据数据类型的不同，可以使用不同的方式进行处理和分析。在后续的章节中，将分别介绍以文本为代表的非结构化数据和以表格为代表的结构化数据的处理与分析的方法。