

面向新型计算模型的系统软件研究新进展与趋势

CCF 系统软件专业委员会

陈海波¹ 廖小飞² 罗英伟³

¹上海交通大学软件学院，上海

²华中科技大学计算机学院，武汉

³北京大学信息科学技术学院，北京

摘要

当前，以大数据、云计算为代表的商业模式推动着新型计算模型的产生。例如，社交计算、推荐系统等推动了大规模图计算的流行与推广，低延迟、高吞吐应用的涌现，以及内存容量的不断增大和新型内存（如非易失内存）的产生催生了内存计算，海量数据的不断生成使得数据密集计算非常重要。新型计算模型也对系统软件提出了新的机遇与挑战。例如，图计算的数据随机访问特征需要系统软件更好地管理并行性与访问局部性，数据密集计算对系统软件的可扩展性等提出了更高要求，内存计算则需要更有效的负载均衡与容错机制。本报告通过对面向各种新型计算模型的系统软件支撑方法进行调研，从图计算、内存计算与数据密集计算三种新型计算模型对近年来国内外在系统软件相关的研究进展进行综述和比较，并对未来的发展趋势进行分析。

关键词：图计算，内存计算，数据密集计算，系统软件

Abstract

Nowadays, business models such as big data and cloud computing are stimulating the emergence of new computing models. For example, social computing and recommendation systems popularized large-scale graph computing; in-memory computing emerges due to the urgent demands of low-latency and high throughput applications, as well as the increasing memory volume and new memory technologies such as non-volatile memory; large-scale data being generated everyday makes data-intensive computing more relevant than ever before. New computing models also raise grand challenges and opportunities to systems software. For instance, the random access characteristics of graph processing demands better management of parallelism and locality by the underlying systems software; data-intensive computing places higher requirements to the scalability of systems software; in-memory computing instead requires better load balance and fault tolerance mechanisms. In this report, we make a thorough survey on the systems software support for the new computing models, i.e. graph computing, in-memory computing and data-intensive computing. Through a comparative study on recent domestic and global advances in systems software, this report further analyzes the research trend for the three models in future.

Keywords: Graph Computing, In-memory Computing, Data-intensive Computing, Systems Software

系统软件通过管理与控制硬件并且为应用软件提供执行环境，在计算机软硬件栈中起到了承上启下的作用，也对各种新型计算模型提供了关键的支撑作用。本报告通过对当前新型计算模型进行调研，选取图计算、内存计算与数据密集计算等作为代表，从应用的特征出发探索系统软件支撑方法的新特征、新需求、新技术以及新的解决方案。通过对比国内外近年来的研究进展，对目前国内研究现状与进展进行总结，并通过分析与对比，对未来可能的研究领域进行展望。

本报告中涉及三项计算模型的关键支撑方法。1) 大数据时代使得对海量数据进行高效准确分析以挖掘潜在价值的需求变得尤为迫切，而许多数据都可以通过图的方式进行存储与处理。在此背景下，面向大规模图计算系统的研究已成为当前并行与分布式系统软件的重要热点之一。相比较为成熟的数据并行处理，图并行系统面临着诸多困难与挑战，如何挖掘图计算特征（如数据关联性、访问局部性与计算迭代性等）对图并行系统的数据划分、并行处理等，是提高图并行计算编程效率与执行性能的关键。由上海交通大学陈海波团队撰写的“图并行计算系统的研究进展与趋势”（陈榕、陈海波等）对国内外图计算方面的系统方法进行了全面的介绍并对未来发展趋势进行探讨。2) 应用对延迟与吞吐等的要求不断提高与内存容量的不断扩大以及新型非易失性内存的出现等催生了内存计算，这对系统软件如何提供低延迟、高吞吐的同时保持系统的高可用性与系统状态的一致性变得非常重要。由华中科技大学廖小飞团队撰写的“内存计算的系统软件支撑方法”（廖小飞等）针对内存计算所面临的挑战，综述了近年来国内外在性能与可用性方面的研究进展，基于分析与对比的结果，进而对其未来研究趋势进行了展望。3) 大数据时代的4V特性也使得数据密集计算变得非常重要，这对传统计算框架的规模提出了更高要求，同时也对如何设计相应的系统软件管理大规模的集群并为用户开发者提供易用接口提出了新的挑战。由上海交通大学陈海波团队与北京大学罗英伟团队联合撰写的“数据密集计算系统的研究进展与趋势”（陈彦哲、陈榕、陈海波，罗英伟等）调研了学术界与产业界关于数据密集计算系统的研究工作并分析了未来的发展趋势。值得注意的是，由于各种计算模型之间在应用形态方面存在一定的重叠，本文描述的三种计算模型尚未存在严格意义上的划分与分类，而是从业界相对较认可的角度进行划分。

1 图并行计算系统的研究进展与趋势

1.1 引言

随着大数据（Big Data）时代的来临，对于海量数据进行高效准确分析以挖掘潜在价值的需求变得尤为迫切。过去十年，面向数据并行（Data-parallel）计算系统的研究一直

是学术界和产业界关注的热点，并建立起以 Hadoop 为核心的开源生态系统。然而，随着数据分析应用场景的拓展和对于数据挖掘精度需求的进一步提高，数据间的关联性成为分析过程中不可忽视的因素。大规模数据及其关联性需要以图数据结构进行描述，并使用机器学习和数据挖掘等算法加以深度分析和挖掘。例如，社交网络使用 Clustering 算法分析用户群落，搜索服务采用 PageRank 算法评估结果相关度，视频网站基于 Collaborative Filtering 算法提供影视推荐等。

然而，由于图并行计算在数据存储和算法行为上的特性造成现有数据并行计算系统（如 Hadoop）无法提供高效的支持，缺乏对数据间关联性的描述手段以及对迭代计算的低效支持可能造成数十倍乃至数百倍的性能损失。面向大规模图计算系统的研究已成为当前并行与分布式处理领域的重要课题之一，产业界与学术界涌现出大量开拓性研究成果。然而，目前仍然处于发展期的图并行计算系统研究尚缺乏针对图数据和算法特征、计算任务负载特性，以及新型硬件架构支持等方面的深入探讨和尝试。本文将首先探讨图并行计算所面临的挑战，并结合对国内外相关研究的介绍和分析，展示当前图并行计算系统的发展现状，进而对图并行计算系统的未来发展趋势进行展望。

本节旨在让读者对大规模图并行计算系统的出现、现状和发展趋势有一个较为全面而翔实的了解和认知。

1.2 图计算特征和技术挑战

相对于以 MapReduce 为代表的数据并行计算，图并行计算具有以下三大特征：

1) **数据关联性**: 对于数据关联性的关注和研究是以机器学习和数据挖掘为代表的图算法的重要特性之一。例如，社交数据中用户间的关注（Follow）关系，影视推荐系统考量的用户与电影间的评论关系，以及主题建模中文本对词语的包含关系等。如图 1a 中，输入数据需要使用图结构的形式存储，以顶点间的边表示数据之间的依赖关系。数据间的关联性将直接影响图并行计算中数据的划分和任务的并行。

2) **访问本地性**: 图结构中数据在计算过程中对于周围数据的访问具有本地性，即主要通过相邻边对邻接顶点进行读写。以图中心度算法 PageRank 为例，每个顶点通过累积邻接顶点的加权值来更新自身的顶点值。如图 1b 中，顶点 1 中经由入边访问邻接顶点 3、4 和 5 完成对自身值的一次更新。访问本地性为隐藏网络通信延时提供了可能，但同时也导致了数据访问缺乏局部性。

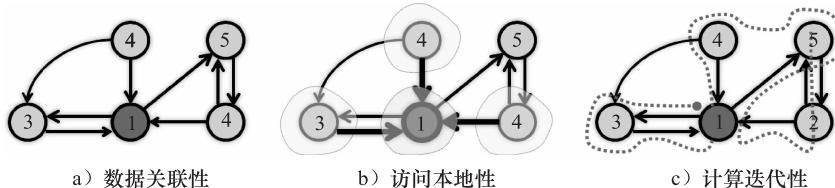


图 1 图并行计算特征

3) **计算迭代性:** 由于数据间的关联性和访问的局部性, 图算法通常被抽象为一个迭代计算过程。顶点对自身值的更新同时会引起邻接顶点的重新计算, 并沿着边进一步扩散直到所有顶点值的变化到达收敛, 即顶点值变化小于某个阈值。如图 1c 中, 顶点 1 值的更新会引起周围邻接顶点的重新计算。迭代计算过程中顶点收敛速度的差异会影响任务调度策略的选择, 以及计算负载的变化。

上述特征决定了现有数据并行计算系统无法有效地支持图并行计算。1) 数据并行计算不考虑数据间的关联性, 计算过程也不提供数据间细粒度的访问支持, 导致大量冗余的数据存储和传递; 2) 数据并行计算只提供了数据块级别的任务调度, 无法满足图算法中顶点级别的细粒度调度需求; 3) 传统数据并行计算并未考虑迭代计算过程, 使用存储持久层在任务间传递数据会造成极大的 I/O 通信开销; 4) 数据并行计算采用的负载均衡和错误恢复等技术都未考虑数据间的依赖关系, 不能满足图并行计算的需要。

因此, 正如中国计算机学会大数据专家委员会在 2013 年发布的“中国大数据技术与产业发展白皮书”^[1]中指出, 对于复杂数据关系的计算任务需要研究和使用图数据计算模式。图并行计算系统的技术挑战主要来自以下几方面:

1) **数据驱动计算:** 首先图计算模式通常是数据驱动, 即依据计算过程中数据的变化经由图结构中的边驱动计算任务。不同于传统任务并行 (Task-parallel) 计算模式通过划分子任务, 并由多个进/线程并行执行来挖掘计算的并行。因此, 数据驱动计算需要通过对数据的划分来实现计算并行, 而图结构数据具有细粒度与非规则依赖, 并具有动态变化特性, 造成了图计算的并行性困难。

2) **关联数据划分:** 如何对大规模图结构数据进行高质量 (均衡且低切割) 和高效率的划分是一个尚待解决的技术挑战。尤其随着数据规模的增加, 传统离线式的迭代划分算法 (例如, Metis) 受限于空间和时间负载度已难以胜任划分需求。同时, 图数据本身的不同特征又对图划分问题提出了新的挑战, 例如幂律性 (Power-law) 和二分性 (Bipartite) 等。

3) **访问局部性差:** 由于图结构数据间依赖关系的不规则与非结构化导致数据访问模式的局部性较差, 当前基于时空局部性设计的处理器难以通过硬件数据预取来隐藏访问延时。内存计算仅能在一定程度上缓解局部性问题, 但当前体系结构在缓存和内存设计上的层次化发展趋势将进一步加剧局部性缺失问题对图计算系统造成的性能影响。

4) **计算负载差异:** 不同图算法的计算与访存负载比率差异较大, 例如 Alternating Least Squares (ALS) 和 Loopy Belief Propagation (LBP) 属于典型 CPU 密集型, 而 PageRank 和 SSSP 则属于典型 I/O 密集型。图计算系统需要采用适合的调度策略以获得最佳性能。此外, 输入数据规模、硬件配置等因素, 以及算法执行过程中计算类型的变化都对图计算系统的设计提出了更大的挑战。

5) **数据特征多样:** 相对于传统科学计算使用的规则图结构数据, 以机器学习和数据挖掘算法为代表的大数据分析应用中的输入数据多来自于真实应用场景 (例如, 社交网络和视频网站等)。自然图在结构上趋于多样性。例如, 社交网络中关注图具有幂律性特征, 而视频网站评论数据的二分性特征。如何对具有不同特征的数据和算法提供高效支

持是图计算系统面临的又一技术挑战。

针对上述技术挑战，图并行计算系统的研究主要从编程模型、数据划分、任务调度和通信机制等方面提出不同设计策略并展开研究。

1.3 系统设计策略

当前图并行计算系统在设计策略上主要考量以下四个方面：

1) **编程接口：**当前主流图并行计算系统采用由 Google 公司在 Pregel 系统中提出的“以顶点为中心”(Vertex-centric) 的编程接口^[2]。该接口遵循“Think as a vertex”原则，要求程序员将图算法逻辑抽象成对顶点的计算，并沿边进行数据访问。图 2 给出如何从 PageRank 算法公式抽象得到以

顶点为中心的图并行程序，并总结典型实现包含的三个步骤（收集邻接顶点数据、更新顶点自身数据和激活邻接顶点）。除此之外，研究人员也尝试提出了支持“以边为中心”^[3] 和“以图为中心”^[4] 的编程接口的图并行计算系统，或将“以顶点为中心”的编程接口进一步分解以支持不同粒度的并行需求^[5]。

2) **划分算法：**受限于单机硬件资源规模，对于大规模数据的分析需要使用图划分算法将存储和计算任务分散到集群的各节点或外部存储层（如固态硬盘）。因此，图划分算法的质量和效果将直接影响图计算系统的性能。由于传统离线划分算法难以应用在大规模图数据上，当前图计算系统多采用仅使用部分顶点和边信息进行划分的在线算法。在线划分算法又可分为边切割 (Edge-cut) 策略和顶点切割 (Vertex-cut) 策略。如图 3 所示，边切割策略首先在各分块间均匀分配顶点，并为所有顶点创建边以构建本地子图。顶点切割策略首先在各分块间均匀分配边，并为所有边创建顶点及副本，单个顶点的计算被分散到多个分块。

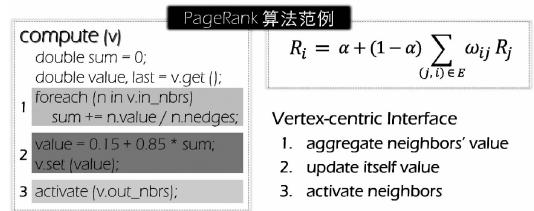


图 2 以顶点为中心的编程接口

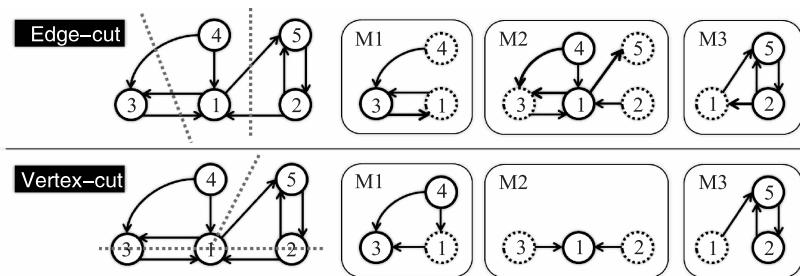


图 3 图划分算法

3) **通信机制：**图并行计算系统主要采用两种通信机制实现顶点间数据的传递，分别是基于推模式 (Push-mode) 的消息传递机制 (Message Passing) 和基于拉模式 (Pull-

mode) 的分布式共享内存机制 (Distributed Shared Memory)。如图 4 所示, 在消息传递机制中顶点需要主动向邻接顶点推送顶点信息, 而在分布式共享内存机制中顶点仅需要与自己的副本保持同步, 由邻接顶点主动由本地副本中获取顶点信息。当前图并行计算系统同时支持两种机制以更高效地实现基于不同访问模式的图算法, 如 PageRank 算法在拉模式通信下可避免冗余的计算和通信开销, 而 SSSP 算法在推模式通信下可消除不必要的数据访问。

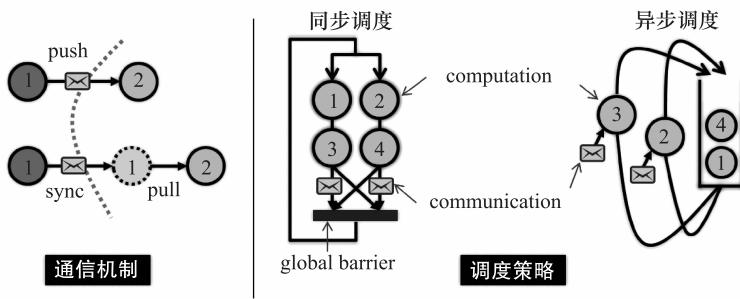


图 4 通信机制和调度策略

4) 调度策略: 典型图算法被抽象为一个迭代计算过程, 通常有两种对于顶点的计算和通信的调度策略: 同步和异步模式。如图 4 所示, 同步调度引擎使用一系列运行在多个工作进程上的迭代轮描述计算过程, 每一轮迭代由三个阶段组成: 顶点计算 (并发更新顶点)、消息通信 (交换顶点消息) 和全局同步 (确保进度一致)。异步调度引擎则使用分布式调度队列在多个工作进程上维护活动顶点, 并发的调度顶点执行计算和通信操作而无需全局同步。同步与异步调度的最大差别在于顶点计算时前者使用上一轮结束时的邻接顶点状态而后者使用当前最新的邻接顶点状态。

1.4 国外研究现状

图并行计算系统最先在国外产业界兴起并成为当前学术界的研究热点之一。图计算系统及相关技术不仅来自于国外高校和开源社区, 同样也得到主流 IT 企业的大力支持。下面主要介绍国外在图并行计算系统领域的一些研究动态。

Google 公司由于自身业务的需求率先开展了对于大规模图并行计算系统的研究, 提出基于经典 BSP (Bulk Synchronous Parallel) 模型^[120]的 Pregel^[2]系统, 其中 “Think as a vertex” 理念和 “以顶点为中心” (Vertex-centric) 的编程接口被后续图计算系统广泛采用。Pregel 系统遵循类 BSP 模型设计采用同步调度引擎和纯消息传递机制, 图算法的执行具有确定性 (Deterministic) 特征易于在分布式环境中编程和调试。Pregel 系统同时提出消息合并和全局聚类接口等扩展优化系统性能和增强应用支持, 并采用基于同步 Checkpoint 技术实现系统容错支持。基于 Pregel 系统在图应用上性能和表现力方面的优势, 开源软件社区随即基于类似思想在以 Hadoop 为核心的大规模数据存储和计算生态环

境中建立了开源图计算系统 Giraph^[6] 和 Hama^[7]，并得到来自 Yahoo 和 Facebook 等公司的技术支持，并将学术界在图计算领域的最新成果应用于开源项目中。

基于纯消息通信机制的 Pregel 系统及其开源实现不能为采用拉模式的图算法（例如，PageRank 等）提供动态计算（Dynamic Computation）支持，即在整个执行过程中所有顶点必须处于活跃状态参与计算并主动向邻接顶点推送消息，无论自身的值变化是否到达收敛，进而造成大量不必要的计算和通信开销。针对上述问题，卡内基梅隆大学 SELECT 实验室的研究人员提出了基于由顶点复制技术实现的分布式共享内存抽象的 GraphLab^[8] 系统。GraphLab 通过增加额外的一轮消息通信提供对动态计算的完整支持。此外，GraphLab 系统采用异步调度策略加速计算的收敛，对于 CPU 密集型的机器学习和数据挖掘算法能获得较大的性能提升，但也造成执行过程不再具有确定性，从而为程序员的编程和调试带来困难。

针对已有图计算系统不能高效划分和处理具有幂律性（Power-law）特征的自然图，例如社交网络中的关注图，GraphLab 的开发人员提出基于顶点切割策略和 GAS 编程模型的 PowerGraph^[5] 系统。PowerGraph 使用顶点切割（Vertex-cut）策略避免对边及边上数据的复制，同时保证了划分后集群中各节点上边数量的平衡。在此基础上，利用 GAS 模型将“以顶点为中心”的图算法操作进一步细分，将原来基于顶点的并行转化为基于边的并行。原本单一顶点上的计算任务被分解到多个节点上，改善了计算并行性和负载均衡。并同时提供了对同步和异步调度策略，以及消息传递和分布式共享内存的支持。

来自斯坦福大学 InfoLab 实验室的研究人员基于类 BSP 模型开发了 GPS^[9] 图计算系统。GPS 系统相对于 Pregel 及开源衍生实现的区别主要体现为：首先支持使用面向领域的高层语言 Green-Marl^[10] 编写图算法，其次支持在计算过程中动态调整图划分迁移顶点提升系统性能，最后针对顶点边分布不均的自然图提出 LALP 数据结构以减少计算过程中的消息通信量。KAUST 大学和 IBM 公司的研究者同样提出了基于动态迁移顶点实现图计算过程的负载均衡的图计算系统 Mizan^[11]。

加州大学伯克利分校 AMPLab 实验室的研究人员在开源内存计算平台 Spark 基础上实现 GraphX^[12] 系统提供对图计算应用的支持。GraphX 系统提出将数据并行计算与图并行计算统一到以表形式存在于内存中的大规模数据上，当需要执行图计算相关操作时，借由对多张表的 Join 操作在内存中按需构建图结构数据。GraphX 系统在图计算方面的性能受限于数据转换操作，但当整个数据分析过程由多个数据并行和图并行计算阶段构成时，能够提供较优的整体性能。

康奈尔大学 BigRedData 实验室的研究人员提出 GRACE^[13] 系统将同步编程模型接口和异步调度引擎相结合，允许程序员使用同步接口和调度策略进行图计算应用的开发和调试，而实际使用时可以选择采用异步调度引擎以获得更好的性能。在此基础上，GRACE 系统进一步针对 I/O 密集型图应用的单次顶点计算的计算量较小会影响系统扩展性问题，提出面向 Block 的计算模型^[14]。通过在分布式环境中增加局部迭代次数的方式减少缓存缺失数，克服内存墙（Memory wall）造成的性能损失。

微软研究院首先在多核平台针对图算法特征和多核平台特点，提出包括图划分策略、

内存顶点排列、批量计算和负载均衡在内的一系列优化技术^[20]。微软研究院还与华盛顿大学合作针对流数据（Streaming）的图计算需求，提出 Kineograph^[22] 系统利用 Epoch Commit 协议将连续计算转化为间隔性的增量计算并利用类 GraphLab 模型进行处理。

基于“Think as a vertex”设计原则的图计算系统具有较佳的编程友好性，但缺乏对图划分信息的了解也导致无法针对算法特征进行有效的优化。例如，Pregel 系统中冗余的消息传递和 GraphLab 系统中为保持计算过程中数据的一致性而引起的巨大调度开销。IBM 的研究者提出“Think as a graph”设计思想并实现 Giraph ++^[4] 图计算系统，通过将图数据在分布式环境下的划分信息暴露给上层图算法获得性能提升。

随着众核平台计算能力和存储容量的不断提升，众核服务器为图并行计算系统的研究提供了另一个舞台。卡内基梅隆大学 SELECT 实验室的研究人员使用 GraphChi^[15] 在一台 Mac Mini 上借助固态硬盘扩展存储能力实现了对 Billion 级图数据的计算任务。GraphChi 利用并行滑动窗口（Parallel Sliding Window）技术避免图计算过程中对于固态硬盘上数据的乱序访问，最大限度优化图计算过程中的 I/O 性能。卡内基梅隆大学的研究者提出 Ligra^[16] 系统，针对图遍历算法对子图进行频繁操作的特点进行优化，并能通过图数据密集程度选择合适的数据结构和调度策略。由德克萨斯州大学奥斯汀分校 ISS 实验室开发的基于 Galois 平台的图计算系统^[17] 通过使用更为通用的编程接口使得图算法的实现能够获得数倍性能提升，并针对具有 High Diameter 特征的图数据通过允许图算法依据自身特征定义优先级调度算法加速系统性能。瑞士洛桑联邦理工学院 LABOS 实验室的研究者提出基于“以边为中心”（Edge-centric）编程模型的 X-Stream^[3] 系统。针对基于“以顶点为中心”的图计算系统在访问邻接顶点时局部性较差的问题，在单机环境下采用遍历边执行计算的方式挖掘图计算过程的局部性，并通过在迭代间对消息进行排序进一步挖掘通信阶段的局部性。

学术界同样尝试将 GPU 等加速处理器平台作为图并行计算系统潜在的理想执行环境。新加坡南洋理工大学的研究者在 GPU 平台上构建通用图计算系统 Medusa^[18]，采用“以顶点为中心”的编程接口，并基于 GPU 平台特性利用图划分技术减少 PCI-e 上的通信开销。此外，通过细化用户接口实现在边粒度上的计算和同行并行。加拿大不列颠哥伦比亚大学 NetSysLab 实验室的研究者在由 CPU 和 GPU 构成的混合众核平台上实现了 TOTEM^[19] 图计算系统，并利用基于顶点邻接边数的图划分策略在 CPU 和 GPU 间分配计算负载提高系统性能。

1.5 国内研究现状

图并行计算系统的研究在国内尚处于起步阶段，但越来越多的高校和科研机构的研究人员开始投身该领域研究，并取得了一定成果。

微软亚洲研究院在国内较早开展图并行计算相关研究，设计并实现了分布式图存储和计算系统 Trinity^[21]，能够同时支持联机图查询和离线图分析两种计算任务。针对面向随时间变化的一系列图（Temporal Graph）的计算任务，微软亚洲研究院与清华大学合作

设计并实现了 Chronos^[23] 系统，通过挖掘多次计算间对顶点和边数据的访问局部性提升系统性能。

国防科技大学开展了面向大规模图并行计算的并行系统结构相关研究，在系统体系结构和大内存方面提出一些新的设想，并在天河二号高性能计算机系统上进行了初步优化尝试，并在体系结构和算法方面尚有大量优化提升空间。

由中科院计算所研究人员构建的 GRE (Graph Runtime Engine)^[24] 图计算系统，采用 Scatter-Combine 编程模型和基于 Agent-Graph 的图划分算法降低图计算过程中的通信开销；针对图遍历算法中对于细粒度锁的需求特点，提出 vLock^[25] 技术利用锁的虚拟映射优化锁操作的局部性；提出一种基于图数据连通性的访问局部性建模方法，并对图遍历算法在随机和非随机图下的访问局部性进行了评测^[26]。中科院软件所的研究人员还对图遍历的局部性进行了建模并对最短路径遍历进行了优化^[112,113]。

东北大学的研究人员对云计算环境下的大规模图数据处理技术进行了全方位的调研，并探讨了未来的研究方向^[27]。Maite^[28] 异步图处理框架采用基于差异累积的（Delta-based Accumulative）迭代计算避免迭代间的冗余计算。BHP^[29] 图划分算法利用虚拟桶实现面向 BSP 模型的负载均衡 Hash 图划分算法，相对于启发式图划分算法有效降低了图加载过程中数据的迁移量。

北京大学的研究人员提出图划分感知的图计算引擎 PAGE^[30]，通过联机监控执行过程中的系统状态调整计算引擎的并行策略；Seraph^[31] 图计算系统采用数据与计算模型分离的方式和写时拷贝机制优化面向同一图数据的多个并发计算任务，并采用延后镜像协议为在不同时间提交的备份请求生成一致的内存图镜像降低容错开销。针对实时大规模图计算需求，提出 IncGraph^[32] 系统采用一系列的增量局部图计算任务兼顾实时性和高效性，避免重复计算带来的性能开销。

上海交通大学 IPADS 实验室开展了多项图并行计算相关研究。基于对典型图算法行为的分析，提出基于分布式不可变视图抽象的 Cyclops^[33] 分布式图计算系统，在提供对动态计算支持的情况下，最大限度地降低计算过程中的通信开销，并针对多核集群进行优化。在此基础上，针对图计算特征提出基于顶点复制（Replication-based）的细粒度容错技术^[34]，相对于传统基于检查点备份（Checkpoint-based）的粗粒度容错技术显著降低了性能损失并缩短了恢复时间；针对同步引擎收敛速度慢而异步引擎调度开销大的不足，提出基于同步计算调度和异步数据传输的复合计算引擎^[35]；基于图数据不同特征，研究人员提出差异化图划分和图计算思想，针对具有幂律性特征的自然图使用混合图划分算法和计算引擎 PowerLyra^[36]，在分布式环境中同时兼顾并行性与局部性。针对具有二分性的图数据和算法，提出具有偏向性的图划分算法 BiGraph^[37]，在有效降低内存和通信开销同时，挖掘图加载过程的数据亲和性显著减少数据在集群内节点间的传递。

1.6 国内外研究进展比较

由表 1 列出的国内外学术界与产业界在图并行计算领域的研究成果可以看出，国内

对于该领域的相关研究虽然起步稍晚但发展较快，但相关成果在创新性和影响力方面与国际前沿仍有差距。当前国内在该领域的研究集中于对图并行计算平台的构建和图划分算法优化。此外，国内的研究主体主要是高校和科研院所，而国外的研究主体除学术界之外，产业界如谷歌、脸书、微软、IBM 等也积极参与并有公开学术成果发表或向开源社区做出贡献。国内产业界对图并行计算系统具有很强的需求和发展基础，例如，以微信和微博为代表的社交网站和以淘宝为代表的电商网站等均拥有数亿用户数和单天上百亿的交易和信息量等待数据掘宝。但是目前我们比较少发现他们在该领域公开发表的学术成果或开源系统。

表 1 国内外图并行计算研究对比

		2010 年	2011 年	2012 年	2013 年	2014 年
国外	学术界	GraphLab ^[8]		PowerGraph ^[5] , GraphChi ^[15]	GPS ^[9] , GRACE ^[13] , Ligra ^[16] , X-stream ^[3] , Galois ^[17]	Giraph ++ ^[4] , Medusa ^[18]
	产业界	Pregel ^[2]	Hama ^[7]	Giraph ^[6] , Kineograph ^[22]	Mizan ^[11] , Trinity ^[21]	GraphX ^[12]
国内	学术界				IncGraph ^[32] , Cyclops ^[33] , vLock ^[25] , PAGE ^[30] , Maiter ^[28]	Seraph ^[31] , Chronos ^[23] , Imitator ^[34] , BiGraph ^[37] , PowerLyra ^[36] , GRE ^[24]
	产业界					

1.7 发展趋势与展望

国内外图并行计算系统的研究正步入快速发展时期，通用性分布式或众核图计算平台已得到充分研究，以 Apache Giraph、Graphlab 和 GraphX 为代表的开源平台已开始进入商用。未来的图并行计算系统研究将朝着深入挖掘图数据特征、图算法特征和计算平台特征的方向进一步快速发展。具体来说，其未来发展有如下几方面：

1) **图数据特征**：当前图计算系统研究主要以通用平台为主，尚缺少针对图数据特征优化图计算系统的研究。已有研究仅考虑了幂律性和二分性等少量特征，随着图计算应用领域的拓展，更多具有不同特征和属性的图数据将进入研究者的视野。如何依据图数据特征优化系统的存储和计算将成为之后研究的一个重要方向。

2) **图算法特征**：能够被抽象为基于图结构数据的计算的现实应用并不仅限于当前主要研究的图结构分析类算法（例如，SSSP 和 BFS 等）和机器学习与数据挖掘类算法（例如，ALS 和 SDG 等）。计算机视觉、图像识别、网络模拟等众多领域的应用已被移植到图计算平台上，但图计算系统的编程接口、数据划分和调度引擎等方面均有待开展针对性研究。

3) **硬件平台特征**：随着硬件平台在集成度和种类等方面的发展，图计算系统将获得更广阔的舞台。Xeon Phi 等加速处理器和各类混合处理器架构等都可能成为图计算系统

潜在的理想平台，而网络和存储硬件上的发展（例如，InfiniBand 和 PCM 等）同样对图计算系统的设计带来挑战和机遇。如何利用硬件特性提升图计算系统的性能将会是该领域急需解决的研究课题。

1.8 小结

随着大数据时代对数据分析要求呈现出强关联、高精度和低延迟的需求，大规模图并行计算系统已成为学术界和产业界的研究热点和重点。本文旨在让读者对图并行计算系统的出现、现状和发展趋势有个全面、详细的了解与认知。本文首先介绍了图并行计算特征以及由此带来的技术挑战，进而总结了当前图并行计算系统的主要设计策略，综述了近年来国内外关于图并行计算系统的研究现状。并在此基础上，展望了图计算系统研究的未来发展方向及研究趋势。

2 内存计算的系统软件支撑方法

2.1 引言

大数据带来了 4V 的挑战：规模（Volume），数据量越来越大，从万亿字节（TB）级到千万亿字节（PB）级甚至到十万亿亿字节（ZB）级别；种类（Variety），数据种类繁多，既包括传统的结构化数据又包括诸如文本、视频、图片和音频等非结构化数据，而且非结构化数据的比重在快速增加；价值（Value），数据价值密度低，难以进行预测分析、运营智能、决策支持等计算；速度（Velocity），大数据处理的速度问题愈发突出，时效性难以保证。总体来看，大数据处理的挑战实质上是由信息化设施的处理能力与数据处理的问题规模之间的矛盾引起的。大数据所表现出的增量速度快、时间局部性低等特点，客观上加剧了矛盾的演化，使得以计算为中心的传统模式面临着内存容量有限、输入/输出（I/O）压力大、缓存命中率低、数据处理的总体性能低等诸多挑战，难以取得性能、能耗与成本的最佳平衡，使得目前的计算机系统无法有效处理 PB 级以上的大数据。

近几年，通过对数据组织管理和编程模型进行革新，以内存优先为原则的传统大内存计算方式被提出来，并且显示其能很大提升大数据的处理性能。然而在传统大内存架构系统中，大数据被组织并存储在传统大内存中，系统通过对被存储在内存中的大数据集进行实时查询与分析实现对复杂数据的处理，但大数据集仍需从外存加载，中间计算结果有时还需在外存存储，数据在内存和外存间可能存在频繁交换，而最后的计算结果还需存储在外存，由于内存和外存之间的 I/O 性能并不匹配，数据 I/O 瓶颈仍是这种计

算方式需要解决的重要问题。

因此，整个 IT 架构的革命性重构势在必行。随着 RRAM、FeRAM 和 PCM 等新型非易失性存储器件的出现和成本的不断走低，客观上为设计以数据为中心的大数据处理模式（即内存计算模式）创造了机会。它将新型存储级内存（Storage Class Memory, SCM）器件设计成为新内存体系的一部分，而非作为虚拟内存交换区域的外存补充，计算不仅存在于传统的内存上，也在新型存储级内存上发生。

基于新型存储级内存和传统 DRAM 设计新型混合内存体系，可以在保持成本和能耗优势的前提下大幅提升内存容量，从而避免传统计算设施上内存-磁盘访问模式中 I/O 能力受限的问题，使计算不仅可以在 DRAM 内存上进行，也可以在 SCM 上进行，这对传统的以计算为中心的设计模式提出重大挑战，为大数据处理提供一种基于混合内存架构的以数据为中心的处理模式，从而大幅度提升大数据处理的时效性。这种以新型非易失型存储设备为基础构建混合内存体系以加速计算的模式，被称为内存计算。本文将介绍国内外对内存计算的系统软件支撑方法的研究现状和最新成果，并进行对比分析。

2.2 国际研究现状

国际上针对内存计算的系统软件支撑方法的研究主要集中在内存数据管理和全内存数据并行处理两个方面。同时，由于异构内存能够很好地解决内存计算的性能、能耗与成本的问题，因此它最近成为内存计算的研究热点，在此节我们也会进行相关研究的介绍。

2.2.1 内存数据管理

在内存数据管理方面，目前已经有了很多研究，其中包括分布式内存缓存管理和内存数据库，并提出了很多典型的系统。

(1) Memcached

Memcached^[38]是一种高性能的分布式内存缓存服务器，用以提高 Web 应用扩展性。许多 Web 应用都将数据保存到 RDBMS (Relational DataBase Management System) 中，应用服务器从中读取数据并在浏览器中显示。但随着数据量的增大、访问的集中，就会出现 RDBMS 的负担加重、数据库响应恶化、网站显示延迟等重大影响。Memcached 通过缓存 DBMS 查询结果，减少 DBMS 访问次数，以提高动态分布式应用的速度和可扩展性。

在存储管理方面，Memcache 采用名为 Slab Allocator 的机制，将分配的内存分割成特定长度的块来进行内存管理。该机制解决了传统内存的分配对所有记录简单地进行 malloc 和 free 而导致内存碎片、加重操作系统内存管理器负担的问题。但是，由于分配的是特定长度的内存，因此 Memcached 也可能无法有效利用分配的内存。Memcached 采用 lazy expiration 回收使用超时的记录空间，在内存空间不足时使用 Least Recently Used (LRU) 机制来替换内存空间。

(2) Redis

Redis^[116]是一种非关系型内存数据库。类似 Memcached，Redis 采用的键值存储，可

以达到很好的性能。Redis 支持两种方法自动向磁盘写数据，并且能支持五种数据结构（String、List、Set、Hash、Sorted Set），这些特性使得 Redis 既可以用作主要存储数据库又可以用作其他存储系统的辅助数据库。

当服务器关机时，Redis 提供快照和 AOF（Append-Only File）两种将内存数据写入磁盘方法，以保证数据的持久性。AOF 或快照可以保证系统重启或崩溃时数据的恢复，但是随着负载的增加，数据的完整性要求变得突出。Redis 采用主/从复制（Master/Slave Replication）技术解决此问题。在需要写临时数据或均衡读请求时，可以用从服务器保存数据集。当用户向主服务器写操作时，从服务器可以实时地接收主服务器的数据拷贝，而且用户可以随机连接一个从服务器进行读操作，以便减轻主服务器的负载。

Redis 通过减少内存的策略来提高快照的创建/载入、AOF 的载入/重写、从服务器的同步效率以及提高 Redis 内存存储率以减少额外的硬件成本。内存的分配，采用简单的 zmalloc（内部实现是通过 malloc）实现，会有内存碎片产生。Redis 采用短数据结构、Sharded 结构以及打包比特和字节的方法减少内存碎片的产生。Redis 依赖客户端来实现分布式读写。主从复制时，每次从节点重新连接主节点都要依赖整个快照，无增量复制，从而影响性能和效率问题。因此，在 Redis 中，单点问题比较复杂，不支持自动 sharding，需要依赖程序设定。

（3）RAMCloud

RAMCloud^[40]是由斯坦福大学的 John Ousterhout 在 2009 年提出的基于内存的存储管理技术。与基于磁盘的存储系统相比，它至少能提供 100 ~ 1 000 倍的带宽和 100 ~ 1 000 倍的低延迟的可靠性和持久性的数据存储系统。相比其他的存储系统（Memcached、Redis 等）而言，RAMCloud 的不同点主要在于所有的信息都保存在 DRAM 中，磁盘仅仅用作备份；其次，RAMCloud 能自动扩展支持成千的存储服务器，应用程序只是看到一个单独的存储系统。

RAMCloud 采用日志结构（log-structured）的方法对内存进行管理，可以达到 80% ~ 90% 成的内存利用率，并且提供了良好的性能服务；在内存和备份磁盘上的信息采用统一的日志结构机制组织；用两层清理的策略，在高效的内存利用率的情况下，保持磁盘的带宽和性能 6 倍的提升；用多线程隐藏内存清理产生的开销，和正常的操作并发执行。

RAMCloud 的优势在于满足高吞吐量的需求，代价在于需要每比特很高的价格和能耗，因此，RAMCloud 不适合应用需求相对低访问延迟和成本较低的存储环境。此外，在存在大量数据复制的数据中心环境中，RAMCloud 的优势就不明显了，因为数据的传输延迟是主要性能影响因数，而且数据的复制将会导致 RAMCloud 很难做到强一致性的需求。

从国内外相关研究中可以看到，现有并行处理系统内存管理方面的研究仍然主要是基于传统存储架构展开的，即主存加磁盘的存储结构。虽然 RAMcloud、Memcached 将数据缓存在内存中，但仍然需要使用磁盘作为数据备份。并且，这些系统所实现的内存管理仍然是面向传统的内存结构，没有考虑内存计算中内存体系的非易失性和混合异构特性。因此，我们还需要深入研究适应于混合内存体系结构、面向大内存计算的内存管理技术。

2.2.2 全内存数据并行处理

当前，随着大数据时代的到来，支持全内存并行处理的运行系统已成为学术界和工业界研究和开发的热点，出现了一批具有代表性的全内存数据并行处理系统。

(1) Spark 内存计算系统

Spark^[41]是加州大学伯克利分校 AMP Lab 所开源的类 Hadoop MapReduce 的并行处理运行软件平台，也称并行计算框架。Spark 系统允许用户将交互式分析和迭代计算作业频繁访问的数据缓存在内存中。系统借鉴函数式编程思想设计了简洁优雅的内存计算编程模型 RDD，简化了多阶段作业的流程跟踪、任务重执行和周期性检查点机制的实现。Spark 是基于 MapReduce 算法实现的分布式计算框架，拥有 Hadoop MapReduce 所具有的优点；但它不同于 MapReduce 的是，Job 中间输出和结果可以保存在内存中，从而不再需要读写 HDFS，因此，Spark 能更好地适用于数据挖掘与机器学习等需要迭代的 MapReduce 算法。

Spark 提供了多种数据集操作类型，给开发上层应用的用户提供了方便。各个处理节点之间的通信方式不再像 Hadoop 那样就是唯一的 Data Shuffle 一种模式。用户可以命名、物化、控制中间结果的存储、分区等，因此更为灵活。RDD（Resilient Distributed Dataset）是 Spark 的最基本抽象，是对分布式内存的抽象使用，实现了以操作本地集合的方式来操作分布式数据集的抽象实现。RDD 是 Spark 的最核心抽象，它在集群节点上构建分布式的不可变内存数据集，并提供粗粒度的转换操作由原内存数据集创建新的内存数据集（如 map、filter、join 等）。当出现数据丢失时，能够依据保存的“粗粒度转换操作”自动重建丢失数据达到容错效果。RDD 可以将数据缓存在内存中，每次对 RDD 数据集操作得到的新的 RDD 数据集仍然能够存放到内存中供下一个操作直接使用，从而省去了 MapReduce 类系统中的大量磁盘 I/O 操作。RDD 具有可重构、不变性、分区局部性以及可序列化等特性。

(2) Phoenix 与 Ostrich

Phoenix^[42]是斯坦福大学开发的基于多核/多处理器、共享内存的 MapReduce 实现。它的目标是在多核平台上，使程序执行得更高效，而且使程序员不必关心并发的管理。Phoenix 由一组对程序应用开发者开放的简单 API 和一个高效的运行时系统组成。运行时系统处理程序的并发、资源管理和错误修复，它的实现建立在 P-thread 之上，也可以很方便地移植到其他的共享内存线程库上。上海交通大学并行与分布式系统研究所提出了分块式 MapReduce^[50,51]，通过将 MapReduce 任务切小以适合缓存的大小，并通过软件流水线提高 CPU 利用率，从而显著地提高 MapReduce 任务的访存局部性与应用性能。

(3) 基于内存的流数据处理系统

GridGain^[43]是一个开源的网格计算框架，专注于提供平行计算能力，能够与 JBoss 和 Spring 相集成。它不仅可以通过网格化集成计算能力，而且支持将内存作为数据的主要存储地，从而可以推动企业向基于内存的应用架构转型。S4^[114]由 Yahoo 公司开发，2011 年将其加入 Apache 托管。S4 是一个分布式流计算平台，它有良好的可扩展性，具有部分

容错能力，能够支持插件并且较为通用。应用程序员在 S4 平台上能够敏捷开发流数据处理的相关应用。Storm^[45] 是一个分布式的、容错的实时计算系统，遵循 Eclipse Public License 1.0，可以方便地在一个计算机集群中编写与扩展复杂的实时计算。Storm 每秒可处理数以百万计的消息，而且开发者可以使用任意编程语言来做开发。Storm 自 2011 年发布以来，凭借其优良的实时流计算框架设计及开源特性，已经应用到许多大型互联网公司的实际项目中，如淘宝和阿里巴巴许多业务（如业务监控、广告推荐、买家实时数据分析等业务场景）都需要实时流计算的支撑。

尽管研究人员在全内存数据并行处理方面开展了大量的工作，但目前并未系统地研究对混合内存体系结构的区分，致使任务划分、作业并发扩展等方面无法充分发挥各类体系结构的优势，难以发挥大内存计算的优势，难以保障大数据并行计算的时效性。目前的研究对数据布局的优化考虑较少。尤其针对混合异构内存系统的数据布局研究较少。研究适应于混合内存体系结构、面向多种应用特征的分布式并行系统支撑技术，对充分利用内存计算的优势、满足高时效大数据处理需求至关重要。

（4）基于事务内存的并行处理系统

在传统大内存架构系统中，最近也有不少工作进行了基于事务内存的全内存数据并行处理研究。

K. Manassiev 等人^[46] 实现了在分布式共享内存系统（Distributed Shared Memory, DSM）中基于内存页的多版本控制协议。使用这种对共享数据的版本控制协议，能实现多线程访问的同步，此实现还支持 MySQL 的数据库事务操作。集群中的每个节点任何时候都维持着一份全部共享数据的物理拷贝。对于只读事务来说，所有的操作都可以在本地完成，因此具有较高的效率；而需要更新数据的事务，在提交前需要将修改操作通知给其他所有节点，并等待其他节点的回复消息，并且获得一个全局的令牌，以此实现事务的串行。因此，这个协议的缓存模型也是基于缓存全局一致性的，不能支持大规模集群，支持的处理器个数仅 10 个左右。

由 Christos Kotselidis 等人^[47] 提出的 TCC 协议中，事务在提交时使用了懒惰的有效性检测机制。每个准备提交的事务会进入仲裁阶段，在仲裁阶段中，事务将广播它的读写集。其他的并发事务在接收到广播之后，将自身的读写集与之进行比较，发生冲突的事务将被撤销，以保证其他事务的有效性。TCC 协议的实现使用了一种标签的机制，标签是一个全局的序列号。每个提交的事务在广播读写集之前，需要向主节点请求一个标签。在 TCC 协议的保证下，事务得到了串行化，因此保证了共享数据的一致性。但是由于标签机制的存在，给 TCC 协议带来了一个瓶颈。每当需要提交一个事务时，都需要和主节点通信去获得一个标签，这样也给事务的提交增加了通信延时。

Cluster-STM 系统^[48] 是基于 GasNet 通信库开发的支持大规模集群的软件事务内存系统，它使用 GasNet 来实现远程过程调用。Cluster-STM 支持全局统一地址和弱的缓存一致性，该系统就多种不同的事务内存协议的效率进行了比较，但是此系统只是一个尝试开发原型系统的项目，例如，它支持的任务个数不能超过处理器的数目，不能支持动态创建线程，不支持容错等。

从上述讨论可见，目前在分布式事务内存领域的研究大多都建立在现有的分布式共享内存（DSM）系统或者 UPC 这类基于 PGAS 模型的分布式编程框架之上。由于 DSM 等分布式共享内存系统或编程框架本身都有一套分布式共享数据一致性的维护机制，而建立在它们之上的分布式事务内存系统也提供了一套更高层次的数据一致性维护机制和并发控制机制，其中就会产生冗余的通信开销。而共享数据的分布策略也很难完全适应事务内存本身的需求。

(5) 内存数据库系统

内存容量与处理器核数目的增长也为在内存中进行事务处理带来了新的机遇。

麻省理工学院人工智能与计算机科学实验室的研究者提出了对事务的 TID 进行无锁化并行，并且采用乐观一致性事务提交协议的内存数据库系统 Silo^[44]，从而在多核平台上获取了较好的性能可扩展性。上海交通大学并行与分布式系统研究所的研究人员通过对 Kyoto Cabinet 等内存数据库在多核平台上的性能分析发现，读写锁是制约其在多核平台上性能可扩展性的一个关键因素。为此，他们设计与实现了新型的被动读写锁^[66]，消除了事务在提交的时候对缓存的竞争。此外，针对当前内存数据库同步复杂、可扩展性不高等问题，他们还设计与实现了基于事务内存的键值存储^[118]，并基于此设计与实现了基于事务内存的事务提交协议^[62]，充分利用事务内存的特点，从而在获得较好性能的同时显著地降低了实现复杂度。

2.2.3 异构内存管理

大数据背景下，急需处理的数据量越来越大，对数据处理时效的要求也越来越高，对内存容量和数据 I/O 速度需求也越来越迫切。新兴存储技术和器件的出现，为满足内存容量需求和 I/O 速度需求提供了硬件支撑。然而，这些器件都有着不同于 DRAM 的特性。目前人们倾向于将这些新器件与 DRAM 一起构成异构混合内存以实现增加容量、提升性能的目标。很多学者利用 SCM 与 DRAM 共同构建混合内存系统并对其管理机制进行了研究。

(1) DRAM 与 SCM 统一内存管理

G. Dhiman^[49]提出了将 PCM 和 DRAM 相结合的 PDRAM 系统^[12]，给出了相应的软硬件管理机制。该混合系统在逻辑上采用 PCM 和 DRAM 并列的架构，将 DRAM 和 PCM 进行统一编址。在硬件设计方面，基于不同的内存管理器分别管理 DRAM 和 PCM 中的数据信息。在软件设计方面，重新设计了操作系统内存管理器，通过页面的交换和迁移来进行耗损程度的管理。

T. J. Ham 等人^[118]提出了一套针对 PCM 的异构混合硬件架构，给出了一种新的内存控制方案，改进了架构的设计过程并降低了设计难度。将 CPU 片上内存控制器作为主控制器，DRAM 和 PCM 器件分别由两个片外从属控制器连接和管理。PCM 和 DRAM 器件一起连接在内存总线上，两个从属控制器之间设置了数据通道，保证数据可以在 DRAM 与 PCM 之间的迁移。

一般而言，多种存储管理技术之间的协作是由操作系统的模块负责的，例如虚拟内存模块负责物理内存，文件系统模块负责磁盘等外存。现代操作系统把存储管理划分成多个模块，从而尽可能避免模块内部的更改对其他模块的影响。这种模块间的划分逐渐构建出了一套存储管理框架，使研究人员可以专注于各种局部优化，并通过框架将优化技术应用到各种不同的应用场景。

然而，由于 SCM 技术的新特性使得原有操作系统的设计假设不再成立：内存不再具有统一的访问特性；外存不再是缓慢的块设备；内存管理与文件系统可能作用于同一个物理器件。这意味着原有操作系统的模块划分对于新型非易失存储材料不再适用，目前大部分相关研究分别集中于如何在内存系统中使用 SCM，或以 SCM 为媒介构建存储系统，而关于内存和存储统一访问与管理框架的研究并不多。

Ju-Young Jung^[119]在深入研究了非易失性内存的物理特性、连接方式以及操作系统输入/输出软件栈的层级特性的基础上，提出一种新的系统架构 Memorage，该系统通过扩展现有虚存管理，添加新的内存池，实现对存储和内存的统一管理。但是外存资源具体管理还使用文件系统完成。Shuichi Oikawa^[52]在 NVM 上构建文件系统来管理 NVM 块，通过设定文件块大小与内存页面一致大小，使得文件块可以用作物理内存，并将其地址映射到虚拟内存地址空间的方式，实现 NVM 的内存和存储一致管理。

Mnemosyne^[53]针对 PCM 作为持久化内存的场景，提供了编程接口，使上层可以安全地创建和使用持久化数据结构。这两种方法都提供了机制用于保证数据的一致性更新，以及系统崩溃时失效数据结构的处理。

除此之外，还有许多研究针对这种混合内存架构的其他方面提出了各种各样的改进。例如，如何利用硬件来加速整个过程，如何改进 DRAM 与 SCM 之间的换进换出策略，以及如何在仅使用软件的前提下完成置换条件的判断等。

综合而言，现有研究主要分为两大类：一类是利用不同的内存级设备来构成异构混合内存，另一类是利用可以被直接访问的 SCM 来构建文件系统。此外，还有两种研究：一种是为用户应用提供对象持久化的 API（以 NV-Heaps 为例），另一种是通过从文件系统中借出空闲 Page 给虚拟内存系统来达到动态调整资源分配的目的。

（2）持久区域管理

传统的持久对象系统依靠将对象数据存储在辅助存储器来实现数据的持久性。在过去的几年里，固态硬盘（SSD）改革了存储子系统，非易失内存的出现，使得数据的持久性在内存这一媒介的实现成为可能。因此，对持久内存的探索成为了众多学者研究的热点。

Kryder^[54]列出了 13 种处于不同成熟期的非易失内存技术，并详细分析和比较了这些技术在密度、性能、功耗和成本上的差异，并在此基础上预估了哪些技术在将来（2020 年）最可能成功。Guerra^[55]等提出了一种新型的持久内存的抽象机制（SoftPM），允许 malloc 类型的分配机制，并给出了该机制实现的软件体系架构，同时利用容器机制实现操作的原子性和内存数据的正交持久性。Jung^[56]引入了一种持久内存存储监控器（PRISM），来探索不同持久内存存储设计的权衡，同时可以允许设计人员来检测一个持

久内存存储的底层行为，评估在运行实际工作负载时不同的架构组织的表现。Millard 等^[57]设计和实现了共享式持久对象管理系统（SPOMS），该系统通过内存映射的方式，将持久对象直接映射到应用程序的虚拟地址空间，而不需要存储格式的转换。

除此之外，研究热点也集中在持久内存读写性能、功耗以及耗损均衡的优化策略。例如如何通过写取消和写中断策略来提高性能；如何利用软硬件结合技术将关键性能的页和高频率写页面保存在 DRAM，将非关键性能页面以及低频率写页面保存在 PCM，来提高系统性能；如何通过引入多级优先队列数据结构来对页面的修改频度进行分级，操作系统根据多级队列中页面的修改冷热程度处理 PCM 和 DRAM 的数据读写与页面迁移，可达到较高的执行效率。又比如 Gaurav Dhiman^[58]通过软硬件结合的方式，在内存控制器中实现对物理页面访存信息的管理，在操作系统的内存管理层级通过页面的交换/迁移实现对性能和物理页面耗损的优化。

（3）基于页面热度的能耗管理

基于页面热度的异构内存能耗管理策略的相关研究，主要是页面的基于热度划分技术和页面迁移策略以及降低内存写能耗的其他方法。

1) 基于热度的页面划分技术：Kyu Ho Park^[59]等提出了一种简单的基于页面访问频度的冷热页面划分策略。通过设置一个访问位来统计页面的访问频繁程度来划分冷页面和热页面：连续访问命中的页面为热页面，两次均未被访问的页面为冷页面。该策略并不能反映页面全局热度，具有局限性。Luiz Ramos 等人^[60]提出了一种考虑页面访问频度和时间的 RaPP 策略。该策略利用一个 M 级的 LRU 队列来记录页面的访问信息，页面在生存时间内随着被访问程度不断地提高队列等级。在超过生存时间后降低队列等级。该策略全面地考虑页面在很长时间和最近时间段的访问情况。

2) 页面迁移策略：Kyu Ho Park 等人^[22]提出的策略是简单将划分的冷热页面按照存储器的不同特性进行迁移，将热页面移动到 DRAM 中，PCM 中则放入冷页面。但是由于他们的冷热页面划分特别局限，该策略会产生回迁页面造成额外开销。Dong-Jae Shin 等人^[61]提出了一种避免 Kyu Ho Park 等人提出的策略中出现的页面回迁现象的策略，通过将相近的页面分成组，然后计算每个组的平均权重，根据设定的阈值来确定热页面组和冷页面组。最后，成组的进行迁移来避免出现回迁页面。但是，这样做粒度太大，会迁移一些本不需要迁移的页面。

3) 降低内存写能耗的其他方法：Ping Zhou 等人^[117]提出了减少 PCM 单元写次数的冗余比特位写策略。通过比较待写数据和写入单元的数据，然后仅仅写入比特位发生改变的位，减少了写入的位数，降低了能耗。Moinuddin 等人^[63]提出了改变写入数据的方法来降低能耗。他们利用 PCM 单元写入 0 和写入 1 消耗的能耗差异大的特性，将那些写入 1 的数据改变映射至其他位，仅仅写入低能耗的 0 来达到降低能耗的目的。

4) 可靠性保障机制：Justin Meza^[64]采用软硬件协同技术，利用单一的硬件单元在统一地址空间内，结合 NVM 的大容量和持久特性，利用 load/store 接口进行数据访存，结合日志管理、持久管理和元数据存储和检索等，保证系统的性能、可靠性和安全性。文献^[65]描述了 PCM 的基本错误模型，并指出 PCM 的错误为永久错误，并给出动态重复内

存方法，保证错误发生时 PCM 容量柔性降低。但也给系统带来了较大的开销。同时，在传统的纠错码（ECC）的基础上提出了纠错指针（ECP）技术，记录错误位的位置，并将数值进行修正。

2.3 国内研究进展

在传统计算机的存储访问优化方面，国内提出了并行数据重用模型、面向数据对象 Cache 技术、面向动态分布 Cache 的多种数据分布优化技术、面向共享存储结构的相似页技术等，有效优化了系统性能，提高了并行可扩展性。国防科技大学肖依教授等面向网络环境提出了内存网格 Ramgrid 模型^[115]，通过集成网络计算环境中空闲分布内存资源，构建内存共享缓冲池提高了数据访问效率。清华大学舒继武等开展了针对 PCM 的耐久性及容错机制的相关研究^[67]。在新型固态存储领域，国防科技大学对闪存、RRAM、PCM 等开展了深入的研究，逐步形成了涵盖商业、工业、消费类用户的完整固态存储解决方案体系，孕育出的湖南源科创新公司也是目前业界最专业的固态存储供应商之一。项目组基于固态存储介质，设计实现了一种基于高速 PCI-E 接口的高性能乱序并行闪存存储可扩展系统 P3Stor，充分发掘了系统和芯片内多级并行性，提出了一种两级缓冲日志结构的低延迟混合地址映射机制，可减少大量数据迁移，实现系统损耗均衡，延长闪存使用寿命，峰值性能可以达到 40~50 万 IOPS。这些前期研究的积累将有助于未来继续在这个领域做出更具有开创性的成果。

2.4 国内外研究进展比较

根据上文的相关介绍可以看出，国内对于内存计算的系统软件支撑方法的相关研究还处于初级阶段。虽然国内在存储系统及相关方面的高水平期刊和国际会议上均有不少成果发表，但是这些研究工作的焦点，主要还是集中于传统大内存架构系统中内存管理相关的问题。从这方面的问题和研究成果来看，国内外的研究水平是大致相同的。但是对于内存计算的系统软件支撑方法的研究，国内的研究主体主要是企业，而国外的研究主体主要是高校，并有一些公开学术成果发表。国内工业界在内存计算方面具有很强的实力，例如华为在内存计算方面的研究从底层硬件到上层软件都有很好的工作。但是我们很少发现他们在这方面公开发表学术成果。

2.5 发展趋势与展望

与 DRAM 相比，目前的初期 SCM 产品读写速度尚不均衡，有些还存在写次数有限的局限，将 SCM 与传统 DRAM 相结合，可以发挥二者各自的优势，目前已成为业界的研究热点。基于传统 DRAM 和 SCM 的混合内存体系架构的出现给计算机系统软件的设计带来了挑战。如何打破混合内存介质间的差异性，对混合内存进行统一管理和有效使用，在

混合内存中实现数据的有效组织、可靠存储和高效访问，成为了面向内存计算的混合内存体系架构亟待解决的重要问题。为了发挥新型存储级内存容量巨大、非易失的特性，适应异构存储介质在容量、速度、功耗等方面的差异，充分利用各种计算设备的计算能力，研究如何提供适合内存计算模式的新型编程模型，支持数据的局部性表达、多粒度任务划分以及数据与计算的紧密耦合，研究如何提供分布式环境下的任务调度机制与算法，改进现有大数据处理运行环境的不足等，都是内存计算的系统软件支撑方法研究的发展趋势。

2.6 小结

随着信息技术的发展，被采集、存储和处理的数据量急剧膨胀，数据规模和数据处理能力间的矛盾日益严峻，传统的以计算为中心的系统架构难以应对大数据处理对时效、性能方面的要求，而基于传统大内存架构的系统仍然面对着数据 I/O 的挑战。目前新型非易失性存储器件不断出现，其成本也不断走低，使得内存计算应运而生。然而，新型内存计算架构的提出对全系统的能耗、性能、可靠性和访问便利性等带来了影响和挑战。本文综述了国内外内存计算的系统软件支撑方法，并在此基础上展望了其未来发展方向及研究趋势。

3 数据密集计算系统的研究进展与趋势

3.1 引言

随着互联网和云计算，移动电子商务等信息行业的发展，数据已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。据 IDC（互联网数据中心）预计，到 2015 年全球数据量将接近 8ZB，数据量是 2012 年的近 3 倍，Google 公司网页索引量将超过 500 亿，Facebook 社交网站的超过 8 亿平均活跃用户将每天产生 15TB 数据，Twitter 社交软件的近 1.5 亿活跃用户将每日发送超过 3.4 亿条信息。面对如此浩瀚的数据量，如何计算、分析并挖掘背后潜在的科学或商业价值，已成为大数据时代最为关键的问题。

然而，大数据呈现出 4V（即 Volume、Variety、Velocity 和 Veracity）特性。Volume 体现出数据量正从 TB 级别向 ZB 级别发展；Variety 体现出数据类型从传统的结构化向结构化和非结构化两种数据类型共存过渡；Velocity 体现出对数据处理的响应时间从批处理响应时间到实时的流数据处理响应时间转变；Veracity 则指随着数据来源的多元化造成数据可靠度和数据质量问题。由于面向大数据的计算系统需要追求高并发、高性能读写访问、低功耗等特性，传统高性能计算难以适应上述场景的需求。当下来自开源社区、学

术界和产业界的力量正共同开展面向数据密集计算系统的研究，以构建适合于大数据时代的新型计算模式。

本文旨在让读者全面了解和掌握数据密集计算领域的技术进展和趋势，尤其是在开源领域的发展。

3.2 数据密集计算的技术特征

数据密集计算技术领域分层明显，且各层次之间分工明确，根据解决问题的方面不同，技术生态圈整体可以分为如图 5 所示的三个层结构：底层的数据存储层，中间的计算分析层，上层的集群管理层。

3.2.1 数据存储层

随着大数据时代来临，数据量呈现爆发式增长，且数据类型也由原先的结构化数据向半结构化数据和非结构化数据转变。所谓结构化数据，即行数据，这些数据可以使用二维表结构来逻辑表达实现，通常适于存储于数据库中。与此相对，不适合使用数据库二维逻辑表来表现的数据即称为半结构或非结构化数据，包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频视频信息等。传统的关系型数据库（如 Oracle，Sybase，SQL Server 等）只能满足关系型数据的存储需求，未来在大数据存储层面临的主要挑战是存储快速增长的半结构化和非结构化数据。这类数据的存储系统应当具备高性价比、高可靠性、容量横向扩展和支持分布式计算等特点。

支持数据密集计算的存储层分为两大类别：1) 提供文件类操作支持的分布式文件系统，例如 HDFS^[81] 等；2) 提供数据库类操作支持的分布式数据库，例如 HBase 和 Cassandra 等。面向数据密集型计算的存储层通常构建在有大量普通服务器组成的集群之上，因此在系统设计上更关注横向可扩展性和高容错性。该类存储层能够提供高吞吐量和高并发，但往往牺牲了对传统存储接口的透明支持，需要上层和存储层协作。

3.2.2 计算分析层

随着数据量的不断增加，在存储成为技术难点的同时，如何高效分析海量数据同样面临巨大挑战。在大数据分析领域，传统方式是依靠高性能商用并行数据库，例如 Vertica、Greenplum、Teradata 和 Exadata 等。并行数据库系统（Parallel Database System）是新一代高性能的数据库系统，构建于 MPP 和集群并行计算环境基础上。其技术起源于 20 世纪 70 年代的数据库机（Database Machine）研究，从 20 世纪 90 年代至今，随着处

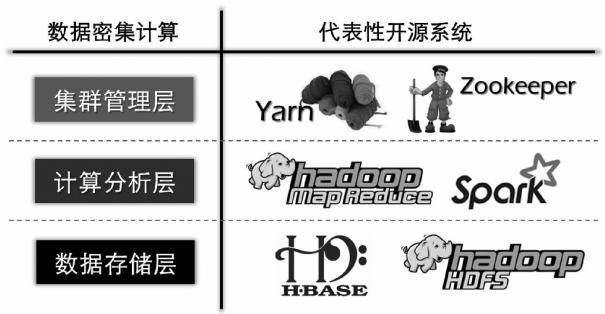


图 5 数据密集计算系统的分层结构和代表开源系统

理器、存储、网络等相关基础技术的发展，并行数据库技术的研究上升到一个新的水平，为大数据分析提供了可行的技术方案。

并行数据库在结构化数据领域提供了高效的数据分析方案，但随着互联网的发展，数据的格式趋于半结构化和非结构化。为了应对这样的趋势，产生了新的数据密集计算技术，最具代表性的便是 Google 公司提出 MapReduce^[68] 数据并行编程模型和分布式计算框架。MapReduce 编程模型借鉴数据函数式编程接口挖掘计算过程中数据间的并行性，实现对大规模集群的计算资源的充分利用。MapReduce 模型仅提供“Map”（映射）和“Reduce”（归约）两个主要编程接口，并向上层程序员隐藏分布式执行带来的各种问题，诸如数据分布、负载均衡和系统容错等。

当前针对半结构化和非结构化数据的数据密集计算技术得到极大关注和广泛应用，但在复杂查询场景下，商业并行数据库仍然具有不可替代的优势。如何在这些领域构建大通用数据密集计算系统已成为当前学术界和产业界的焦点问题之一。

3.2.3 集群管理层

开源项目普遍对系统的易用性重视不足，在数据密集计算领域亦存在这个问题。一个完备的数据密集计算解决方案需要一体化的管理系统，这个系统解决方案涉及：多个独立的开源项目，大量机器构成的集群，不同来源的类型的数据，大量参数的配置、调整和优化，多个作业的部署和运行等。如果依靠手工，没有专门的工具或平台，系统的规模和效率较难得到保证，也影响到系统的推广。国内外众多 IT 企业利用生态系统构建数据密集计算平台，但相互间缺乏必要的沟通和协作，往往各自为战重复开发。目前生态系统的大部分项目均提供一些基本的交互性工具，例如 Shell 和 Web 接口等。但是，普遍功能不全或偏弱，缺少集群的统一配置平台，大量需要手动操作（如增加和删除节点），而配置参数的优化仍然依靠经验。

开源社区已经意识到数据密集计算领域需要统一的集群管理层技术支持。在数据密集计算的集群管理系统中，具有代表性的是 Zookeeper^[69] 和 Apache YARN^[70] 项目。Zookeeper 是一个开源分布式的服务，借鉴 Google 的 Chubby^[71] 系统提供了分布式协作、分布式同步和配置管理等功能。YARN 是一个分布式的资源管理调度系统，用以提高分布式的集群环境下的资源利用率，管理的资源包括内存、I/O、网络、磁盘等。

当前集群管理层已得到各方关注，但主要仍然依靠开源社区的推动。随着集群管理层的不断完善，将极大地简化未来数据密集计算系统的构建。

3.3 国外研究现状

数据密集计算系统的研究由来已久，国外高校学术领域对其特性做了非常深入的研究，与此同时，开源社区和主流 IT 企业也为系统的实现作出了非常多的贡献。下面主要介绍国外在数据密集计算技术领域的一些研究动态。

Google 公司因为业务需求，正在大数据存储领域作出富有开创性的研究，提出 GFS^[72]通过应用层协助分离控制流和数据流，提升集群横向扩展能力，实现对海量数据的高效可靠存储。针对半结构和非结构化数据存储需求，提出 BigTable^[73]分布式类数据库系统，以 GFS 分布式文件系统为基础架构。但不提供传统数据的强事务支持和 SQL 接口，引领了后来的 NoSQL 潮流。这种存储方案被用于 Google 内部多个项目中，如搜索系统的海量数据存储，用户请求的日志管理等，支持起后续 Google 公司的大数据产品（如 Google Analytics、Google Finance 和 Google Earth 等）。Google 近年仍不断在云存储领域开拓，以论文形式公开发表的系统包括增量索引平台 Percolator^[74]，全球级分布式数据库 Spanner^[75]等，在存储响应延时，强一致性保证等提供对数据密集计算的支持。

Amazon 公司基于电子消费类应用行为特征构建高可用和高可伸缩的分布式存储系统 Dynamo^[76]。Dynamo 采用 Key-Value 方式存储数据和提供应用接口，采用数据分块、最终一致性和一致性哈希等技术为上层应用提供 Always-on 体验的同时保持高可用和横向扩展能力。

开源社区建立多个项目在存储层提供对数据密集计算的支持，构建类 GFS 系统的 HDFS 分布式文件系统，采用类似的控制流与数据流分离思想支持存储资源的横向扩展，已成为当前产业界和学术界在密集计算场景下的首选分布式文件系统平台。HBase^[77]和 Cassandra^[78]项目在开源社区填补了类数据库存储空白。HBase 基于 BigTable 的设计思想在 HDFS 的基础上构建，提供较为接近传统数据库的支持。早期由 Facebook 开发的 Cassandra 系统借鉴不同研究成果以 Key-Value 方式构建分布式存储平台，提供更好的扩展性和一致性，后借助开源社区力量不断发展。在数据的可靠性、可用性和分区容错能力（即 CAP^[79]）的折中设计上，采用较为灵活的策略，由用户依据应用需要进行选择。

在计算分析层，受函数式编程的启发，Google 提出 MapReduce^[68] 编程模型，通过将分布式问题与业务逻辑分离的方式构建高效可靠的通用分布式计算框架。利用简单的函数式 Map 和 Reduce 接口向业务程序员隐藏分布式问题，迅速获得成功。该设计思想直接影响了之后的数据密集计算系统的设计。Microsoft 同样提出了自己的数据密集计算平台 Dryad^[80]，相对于 MapReduce 提供的精简接口和静态调度流程，Dryad 提供程序员使用有向无环图（DAG）来描述任务的依赖关系，并提供更为丰富的接口来提升系统表达能力。在 Dryad 平台基础上，Microsoft 进一步构建 DryadLinq^[81] 系统提供对传统 SQL 接口的制动翻译支持。开源社区以 Hadoop^[82] 系统为基础构建面向大规模批量数据的开源计算分析层生态环境，包括用于处理流数据的 Hadoop Streaming^[83]，面向结构化数据的数据仓库 Hive^[84]，以及机器学习系统 Mahout^[85] 等。

以个性化搜索广告为代表的流数据计算需求促使 Yahoo 公司提出基于 Actors 编程模型的分布式流计算系统 S4^[86]。通过动态创建以 Key 和 Value 形式描述的计算单元（Processing Elements），充分利用分布式资源处理大规模流数据。Twitter 公司同样设计了自己的流处理系统 Storm^[87]，弥补了 Hadoop 平台在实时处理方面的局限性。Storm 定义了一批实时计算的原语，利用这些原语在集群环境中动态传递流数据，并由用户定义处

理节点行为简化了并行实时数据处理应用。

HP 公司与加州大学伯克利分校合作基于 R 向量语言提出面向矩阵结构数据的分布式系统 Presto^[88]。该系统支持以矩阵形式描述对于复杂结构数据的分析算法，并提出了针对矩阵结构数据特点的运行时优化策略解决动态数据划分和负载均衡等问题。

加州大学伯克利分校 AMPLab 实验室研发的 Spark^[89] 内存计算系统，迎合了当前硬件发展和对数据处理响应的需求，迅速成为数据密集计算领域的理想平台。围绕 Spark 平台，学术界和开源社区正共同努力构建支持不同需求的数据密集计算生态环境。Spark Streaming^[90] 向上层应用提供类似 Spark 系统的内存计算接口，处理来自多种数据源的实时数据流，并提供有状态的精确语义（Stateful exactly-once semantics）容错。针对传统面向结构化数据的数据库查询请求，提供兼容 Apache Hive 接口的 Spark SQL^[91] 系统。融合数据并行计算和图并行计算的通用框架 GraphX^[92]，以及面向机器学习应用的 MLLib^[93] 等。

Apache Zookeeper^[69] 开源项目是借鉴 Google Chubby^[71] 系统，提供高可靠的分布式协作和同步支持，以及集群配置和管理等功能，为开发者提供易于理解与编程的接口，简化构建分布式系统的过程。

由 Hortonworks 领导的 YARN^[70] 项目设计并实现了一种新的面向 Hadoop 应用的资源管理器，并已被纳入到 Apache 开源社区。YARN 最初是为了修复 Hadoop 实现中在任务调度方面的明显不足，通过将 Job Tracker 的两个主要功能（资源管理和作业调度/监控）分成了两个独立的服务（全局的资源管理和针对每个应用的应用管理），对可伸缩性（支持数万节点和数十万内核的集群）、可靠性和集群利用率进行全面优化。

加州大学伯克利分校 AMPLab 实验室提出的 MESOS^[94] 系统是一款开源集群管理软件和分布式高效调度系统，支持 Hadoop、ElasticSearch、Spark、Storm 和 Kafka 等分布式计算系统，由于其开源性质越来越受到包括 Twitter 和 Facebook 在内的多个大数据服务公司的青睐。

3.4 国内研究现状

数据密集计算系统的研究同样得到国内学术界和产业界关注。

中国计算机学会通信于 2011 年组织撰写《数据密集型计算》专题，邀请多位专家介绍了数据密集型计算面临的挑战。华中科技大学的研究组对数据密集型大规模计算系统从系统结构、数据管理、编程模型、和当时的研究现状进行了深入介绍和分析^[95]。国防科学技术大学的研究人员则从数据密集型计算的系统结构出发，从并行体系结构和芯片设计角度进行分析，提出固态存储介质和计算存储耦合的数据密集计算架构发展方向^[96]。上海交通大学的研究人员从云计算环境下典型数据密集计算系统 MapReduce 和 Dryad 入手就编程模型和自适应资源管理的研究和发展方向进行了深入探讨^[97]。

此外，中科院计算所的研究人员对数据密集计算模式的特点与传统高性能计算进行

了分析和对比，介绍了当时数据密集计算在国际上的发展情况和尚存的主要问题，并进行了总结和展望^[111]。复旦大学的研究者同样针对数据密集型计算存在的问题和面临的挑战进行了不少综述性研究^[98]。

在数据存储层研究上，国内产业界表现活跃。阿里巴巴公司对 Hadoop 进行了深度的重构，开发了自己的分支 ADFS^[99]。ADFS 主要解决的是 HDFS NameNode 在设计上存在单点故障、内存瓶颈，以及集群重启时间过长而期间无法对集群进行写操作等问题。ADFS 原理是将非热点数据保存到外部“数据库”，而非驻于 NameNode 内存中。

OceanBase^[100]是阿里巴巴公司自主研发的可扩展的关系型数据库，实现跨行和跨表的事务支持，能够在数千亿条记录、数百 TB 数据上进行传统 SQL 操作。在阿里巴巴集团下，OceanBase 被应用于多个重要业务的数据存储。海狗（Higo）^[101]是阿里巴巴公司的一个分布式的在线分析查询系统，基于 Hadoop、Lucene、Solr、蓝鲸等开源系统作为实现基础，支持类 SQL 的查询语法。Higo 通过使用索引、列式存储以及内存 cache 等技术优化海量数据在分布式环境下的高效并发查询。

清华大学和阿姆斯特丹自由大学合作研发的 CloudTPS^[102]系统为分布式应用提供存储支持。CloudTPS 在分布式存储系统之上构建独立的中间层来提供本地事务管理（Local Transaction Management）服务，并通过中间层向客户端提供事务性支持。

在计算分析层，国内学术界和研究机构开展了大量工作。针对当前分布式数据流计算系统（例如，MapReduce、Hadoop 和 Dryad 等）难以进行性能调优的问题，英特尔亚太研究院利用分布式代码注入和数据流驱动的性能分析技术，研发了一种可伸缩、轻量级和高可扩展的大规模数据密集计算系统性能调优工具 HiTune^[103]。

上海交通大学并行与分布式系统研究所基于众核硬件特征提出基于分治策略的 Tiled-MapReduce^[104]模型，并提出了输入缓冲区复用、面向非一致性缓存调度机制和软件流水线任务并行机制等面向众核平台的优化策略。基于 Tiled-MapReduce 模型构建的运行时能够提供细粒度的容错支持，并简化了增量计算和在线计算等特殊数据密集计算应用的开发。

针对数据密集计算应用的普及带来的基准测试需求，中科院计算所设计并发布了面向云计算和大数据的基准测试程序集合 ICTBench^[105]，并利用这些基准测试对数据中心负载和大数据应用行为和特征进行了深入分析^[106,107]。

大量数据并行算法可以抽象成基于矩阵的计算。微软亚洲研究院提出 MadLinq^[108]系统，在 Dryad 平台之上构建与面向关系代数处理的 DryadLinq 系统接口兼容的分布式矩阵通用计算平台。MadLinq 在提供矩阵友好的编程接口的同时提供高可伸缩和高效容错支持。

针对数据流的密集计算的低延迟需求，微软亚洲研究院和来自国内多所高校的研究人员共同开发了分布式流处理系统 TimeStream^[109]。TimeStream 通过摒除传统类 MapReduce 模型的批量计算模式，利用一种名为 resilient substitution 的全新抽象迎合流处理过程特性，提供对系统容错和动态重配置的高效支持。

在集群管理层，目前国内公开的研究工作较少，各家公司独立构建了针对大规模集群的管理系统，例如腾讯公司构建的 Torca^[110]是一种分布式集群调度系统。Torca 采用类操作系统的架构设计，负责运行程序，管理 CPU 和内存等资源，其首要设计目标在于提高资源利用率。Torca 可以实现多业务之间的资源共享，也可以实现基于单个机器上的多任务资源共享，并有弹性管理机制，这样就可以根据不同业务的需要，提高资源利用率。此外，Torca 同样提供了系统容错、资源共享与隔离等方面的管理支持。

3.5 国内外研究进展比较

表 2 国内外数据密集计算研究对比

	国外研究	国内研究
数据存储层	GFS ^[72] , BigTable ^[73] , Spanner ^[75] , Dynamo ^[76] , HDFS ^[82] , HBase ^[77] , Cassandra ^[78]	ADFS ^[99] , OceanBase ^[100] , Higo ^[101] , CloudTPS ^[102]
计算分析层	MapReduce ^[68] , Dryad ^[80] , DryadLinq ^[81] , S4 ^[86] , Storm ^[87] , Presto ^[88] , Hadoop ^[82] , Hadoop Streaming ^[83] , Hive ^[84] , Mahout ^[85] , Spark ^[89] , Spark Streaming ^[90] , Spark SQL ^[91] , GraphX ^[92] , MLLib ^[93]	HiTune ^[103] , Tiled-MapReduce ^[104] , DCBench ^[105] , BigDataBench ^[106] , MadLing ^[108] , TimeStream ^[109] ,
集群管理层	Chubby ^[71] , Zookeeper ^[69] , YARN ^[70] , Mesos ^[94] ,	Torca ^[110]

由表 2 列出的国内外数据密集计算领域的研究成果可以看出，国外由于大型 IT 企业对数据密集计算的迫切需求，进而推动相关研究在开源社区、学术界和产业界快速发展，相关成果迅速得到实际应用并不断创新。目前数据密集计算在国外已主要由开源社区负责在功能上逐步完善。然而，大数据带来的变革为数据密集计算研究注入了新的活力，如何支持大数据带来的新需求，如全球级分布式存储和计算、移动接入端和深度机器学习应用等成为当前研究重点。国内对于该领域的相关研究起步相对较晚，但国内产业界对数据密集计算系统的需求极大，数据存储和计算分析能力已成为抑制国内 IT 企业发展的瓶颈。因此，国内产业界已成为数据密集计算研究发展的主力军，多个系统已被应用于实际业务，未来将有机会和国外 IT 企业位于同一前沿展开针对数据密集计算技术的研究竞争。

3.6 发展趋势与展望

随着企业级云计算及虚拟基础设施的不断发展，数据密集计算技术在产业界的应用将越来越广。未来数据密集计算的发展将可能在以下几个方面推进：

1) 平台标准趋于统一：随着类似 Apache YARN、Apache Mesos 等集群管理的出现，典型的计算框架可以共享一套基础系统平台，如离线计算框架 MapReduce、流式计算框架 Storm、内存计算框架 Spark、图计算框架 Giraph 等。

2) **计算模型的多样化**: 根据密集数据的特征以及计算目标的差异, 计算模型将不会局限于单纯的 MapReduce 计算模型一种。如 MapReduce 计算框架用于离线批处理, Spark 计算框架用于基于内存的交互式计算, Storm 计算框架用于流处理计算, 等等。

3) **内存计算占据主导地位**: 随着硬件技术的发展以及对计算响应速度的要求不断提高, 内存计算将在未来数据密集计算领域逐渐呈现出主导地位。此外, 面向复杂互相依赖数据的机器学习和数据挖掘应用的普及同样将推动内存计算的应用和发展。

4) **安全问题更加重要**: 安全问题随着数据密集计算的盛行变得越来越重要。具体来讲, 分为三个方面。首先是计算集群的用户权限管理问题, 大型集群如 Hadoop 等的用户权限管理涉及用户分组管理、集群配置等。其次是存储层的安全策略, 例如 HDFS 中的安全策略、数据节点的身份认证问题。最后是计算层的安全问题, 例如 MapReduce 计算框架中的安全策略、Job 的提交认证等。

3.7 小结

大数据时代引领数据密集计算系统研究成为学术界与产业界的共同热点, 本文旨在让读者对数据密集计算系统的出现、现状和发展趋势有全面、详细的了解与认知。本文首先对数据密集计算系统的构建和分层进行概要介绍, 进而综述了近年来国内外关于数据密集计算系统的研究现状。并在此基础上, 对比国内外研究进展, 展望了数据密集计算领域研究的未来发展方向及研究趋势。

4 结束语

大数据、云计算等驱动的新型计算模型为系统软件研究提供了新的机遇与挑战。本文通过对面向三种新型计算模型(如图计算、内存计算与数据密集计算)的系统软件研究进行调研, 综合分析对比了国内外研究进展, 并对未来进行了展望。从总体上, 国内在这三方面的研究目前紧跟国际前沿, 部分领域接近国际先进水平。然而, 目前的研究主要是基于国外与开源社区现有软件与系统进行改进, 在具有较大原创性的系统软件与平台方面还需要更多的投入。

参考文献

- [1] 大数据专家委员会. 中国大数据技术与产业发展白皮书[J/OL]. 中国计算机学会, 2013.
- [2] G Malewicz, M H Austern, A J Bik, J C Dehnert, I Horn, N Leiser, G Czajkowski. Pregel: a system for large- scale graph processing. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data, pages 135-146[C]. ACM, 2010.

- [3] A Roy, I Mihailovic, W Zwaenepoel. X-stream: edge-centric graph processing using streaming partitions. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, pages 472-488[C]. ACM, 2013.
- [4] Y Tian, A Balmin, S A Corsten, S Tatikonda, J McPherson. From “Think Like a Vertex” to “Think Like a Graph”[C]. In VLDB, 2013.
- [5] J E. Gonzalez, Y Low, H Gu, D Bickson, C Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI), pages 17-30[C]. 2012.
- [6] Apache. The Apache Giraph Project[J/OL]. <http://giraph.apache.org/>.
- [7] Apache. The Apache Hama Project[J/OL]. <http://hama.apache.org/>.
- [8] Y Low, D Bickson, J Gonzalez, C Guestrin, A Kyrola, J M Hellerstein. Distributed GraphLab: a framework for machine learning and data mining in the cloud. Proceedings of the VLDB Endowment, 5(8): 716-727 [C]. 2012.
- [9] S Salihoglu and J Widom. GPS: A graph processing system. Proceedings of the 25th International Conference on Scientific and Statistical Database Management, page 22[C]. ACM, 2013.
- [10] S Hong, S Salihoglu, J Widom and K Olukotun. Simplifying Scalable Graph Processing with a Domain-Specific Language. Code Generation and Optimization (CGO), IEEE/ACM International Symposium on [C]. IEEE, 2014.
- [11] Z. Khayyat, K Awara, A Alonazi, H Jamjoom, D Williams, P Kalnis. Mizan: a system for dynamic load balancing in large-scale graph processing. Proceedings of the 8th ACM European Conference on Computer Systems, pages 169-182[C]. ACM, 2013.
- [12] R S Xin, J E. Gonzalez, M J Franklin, I Stoica. GraphX: A resilient distributed graph system on spark. First International Workshop on Graph Data Management Experiences and Systems, page 2[C]. ACM, 2013.
- [13] G Wang, W Xie, A J Demers, J Gehrke. A synchronous large- scale graph processing made easy. In CIDR, 2013.
- [14] W Xie, G Wang, D Bindel, A J Demers, J Gehrke. Fast Iterative Graph Computation with Block Updates. In VLDB, 2013.
- [15] A Kyrola, G Blelloch, C Guestrin. Graphchi: Large-scale graph computation on just a pc. Proceedings of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI), volume 8, pages 31-46[C]. 2012.
- [16] J Shun and G E. Blelloch. Ligra: a lightweight graph processing framework for shared memory. Proceedings of the 18th ACM SIGPLAN symposium on Principles and practice of parallel programming, pages 135-146 [C]. ACM, 2013.
- [17] D Nguyen, A Lenhardt, K Pingali. A lightweight infrastructure for graph analytics. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, pages 456-471[C]. ACM, 2013.
- [18] J Zhong, B He. Medusa: Simplified Graph Processing on GPUs[J]. TPDS, 2014, 25(6): 1543-1552.
- [19] A Gharaibeh, L B Costa, E Santos-Neto, M Ripeanu. A Yoke of Oxen and a Thousand Chickens for Heavy Lifting Graph Processing[C]. In PACT, 2012.
- [20] V Prabhakaran, M Wu, X Weng, F McSherry, L Zhou, M Haridasan. Managing Large Graphs on Multi-Cores With Graph Awareness[C]. Usenix ATC, 2012.

- [21] B Shao, H Wang, Y Li. Trinity: A distributed graph engine on a memory cloud. Proceedings of the 2013 International Conference on Management of Data, pp. 505-516[C]. 2013.
- [22] R Cheng, J Hong, A Kyrola, Y Miao, X Weng, M Wu, F Yang, L Zhou, F Zhao, E Chen. Kineograph: taking the pulse of a fast- changing and connected world. Proceedings of the 7th ACM european conference on Computer Systems, pages 85-98[C]. ACM, 2012.
- [23] W Han, Y Miao, K Li, M Wu, F Yang, L Zhou, V Prabhakaran, W Chen, E Chen. Chronos: A Graph Engine for Temporal Graph Analysis[C]. In EuroSys, 2014
- [24] J Yan, G Tan, N Sun. GRE: A graph runtime engine for large-scale distributed graph-parallel applications [J/OL]. arXiv preprint arXiv: 1310.5603, 2013. <http://www.ncic.ac.cn/~tgm/GRE/>
- [25] J Yan, G Tan, X Zhang, E Yao, N Sun. vlock: Lock virtualization mechanism for exploiting fine-grained parallelism in graph traversal algorithms. Code Generation and Optimization (CGO), IEEE/ACM International Symposium on, pages 1-10[C]. IEEE, 2013.
- [26] L Yuan, C Ding, D Tefankovic, Y Zhang. Modeling the Locality in Graph Traversals, International Conference on Parallel Processing (ICPP)[C]. 2012.
- [27] 于戈, 谷峪, 鲍玉斌, 王志刚. 云计算环境下的大规模图数据处理技术[J]. 计算机学报, 2011, 34(10): 1753-1768.
- [28] Y Zhang, Q Gao, L Gao, C Wang. Maiter: An asynchronous graph processing framework for delta-based accumulative iterative computation[J]. TPDS, 2013.
- [29] 周爽, 鲍玉斌, 王志刚, 冷芳玲, 于戈, 邓超, 郭磊涛. BHP: 面向 BSP 模型的负载均衡 Hash 图数据划分[J]. 计算机科学与探索, 2014, 8(1): 40-50.
- [30] Y Shao, J Yao, B Cui, L Ma. PAGE: A Partition aware Graph Computation Engine. 22th International Conference on Information and knowledge management (CIKM)[C]. 2013.
- [31] J Xue, Z. Yang, Z. Qu, S Hou and Y Dai. Seraph: an Efficient, Low-cost System for Concurrent Graph Processing. ACM Symposium on High- Performance Parallel and Distributed Computing (HPDC) [C]. Vancouver, Canada, 2014.
- [32] 申林, 薛继龙, 曲直, 杨智, 代亚非. IncGraph: 支持实时计算的大规模增量图处理系统[J]. 计算机科学与探索. 2013.
- [33] R Chen, X Ding, P Wang, H Chen, B Zang and H Guan. Computation and Communication Efficient Graph Processing with Distributed Immutable View. ACM Symposium on High- Performance Parallel and Distributed Computing (HPDC)[C]. Vancouver, Canada, 2014.
- [34] P Wang, K Zhang, R Chen, H Chen, H Guan. Replication-based Fault-tolerance for Large-scale Graph Processing. The 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)[C]. Atlanta, USA, 2014.
- [35] 丁鑫, 陈榕, 陈海波. 分布式图计算框架混合计算模式的研究[J]. 小型微型计算机系统, 2014.
- [36] R Chen, J Shi, Y Chen, H Guan, B Zang, H Chen. PowerLyra: Differentiated Graph Computation and Partitioning on Skewed Graphs [J/OL]. IPADS- TR, 2013. <http://ipads.se.sjtu.edu.cn/projects/powerlyra.html>.
- [37] R Chen, J Shi, B Zang and H Guan. Bipartite-oriented Distributed Graph Partitioning for Big Learning. The 5th Asia-Pacific Workshop on Systems (APSys)[C]. Beijing, China, 2014.
- [38] Memcached, Mar. 2013. <http://www.memcached.org/>.
- [39] Zhaoguo Wang, Hao Qian, Jinyang Li, Haibo Chen. Using Restricted Transactional Memory to Build a

- Scalable In-Memory Database. The European Conference on Computer Systems (EuroSys 2014) [C]. Amsterdam, The Netherlands, 2014.
- [40] Ousterhout, J., Agrawal, P., Erickson, D., Kozyrakis, C., Leverich, J., et al. The case for ramcloud [J]. Communication of the ACM, 2011, 54: 121-130.
- [41] Zaharia, Matei, et al. Spark: cluster computing with working sets. Proceedings of the 2nd USENIX conference on hot topics in cloud computing[C]. 2010.
- [42] Yoo R M, Romano A, Kozyrakis C. Phoenix rebirth: Scalable MapReduce on a large-scale shared-memory system. Proceedings of the 2009 IEEE International Symposium on Workload Characterization[C]. 2009.
- [43] GridGain, Mar. 2013. <http://www.gridgain.org/>.
- [44] S Tu, W Zheng, E. Kohler, B Liskov, S Madden. Speedy Transactions in Multicore In-Memory Databases. Proc. SOSP[C]. 2013.
- [45] Storm, Mar. 2013. <https://storm.incubator.apache.org/>.
- [46] Manassiev, K, Mihailescu, M, Amza, P. Exploiting Distributed Version Concurrency in A Transactional Memory Cluster. In PPoPP'06[C]. 2006.
- [47] Kotselidis, C, Ansari, M, Jarvis, K, Lujan, M, Kirkham, C, Watson, I. Distm: A Software Transactional Memory Framework for Clusters. In ICPP'08[C]. 2008.
- [48] Bocchino, R L, Adve, V S, Chamberlain, B L. Software Transactional Memory for Large Scale Clusters. In PPoPP'08. 2008.
- [49] Dhiman, Gaurav, Raid Ayoub, Tajana Rosing. PDRAM: a hybrid PRAM and DRAM main memory system. Proceedings of the 46th ACM/IEEE Design Automation Conference[C]. 2009.
- [50] Rong Chen and Haibo Chen. Tiled-MapReduce: Efficient and Flexible MapReduce Processing on Multicore with Tiling[J]. ACM Transactions on Architecture and Code Optimization (TACO). Volume 10, Issue 1, Article No. 3. April, 2013.
- [51] Rong Chen, Haibo Chen and Binyu Zang. Tiled MapReduce: Optimizing Resource Usages of Data-parallel Applications on Multicore with Tiling. The Nineteenth International Conference on Parallel Architectures and Compilation Techniques (PACT 2010) [C]. Vienna, Austria, 2010.
- [52] Shuichi Oikawa, Satoshi Miki. File-based Memory Management for Non-Volatile Main Memory. In COMPSAC'13[C]. 2013.
- [53] Volos, Haris and Tack, Andres Jaan and Swift, Michael M Mnemosyne. lightweight persistent memory. Proceedings of the 16th international conference on Architectural support for programming languages and operating systems[C]. 2011.
- [54] Kryder M H, Kim C S. After hard drives—What comes next? [J]. IEEE Transactions on Magnetics, 2009, 45(10): 3406-3413.
- [55] Jorge Guerra, Leonardo Marmol, Daniel Campello, Carlos Crespo, Raju Rangaswami, Jinpeng Wei. Software Persistent Memory. Proceedings of the USENIX Annual Technical Conference[C]. 2012.
- [56] Ju-Young Jung, Sangyeun Cho. PRISM: Zooming in Persistent RAM Storage Behavior. Proceedings of the 2011 IEEE International Symposium on Performance Analysis of Systems and Software[C]. 2011.
- [57] Millard B R, Dasgupta P, Rao S, et al. Run-time support and storage management for memory-mapped persistent objects. Proceedings of the 13th International Conference on Distributed Computing Systems [C]. 1993.
- [58] Gaurav Dhiman, et al. PDRAM: A Hybrid PRAM and DRAM Main Memory System. In DAC'09[C]. 2009.

- [59] Park, Kyu Ho, et al. Efficient memory management of a hierarchical and a hybrid main memory for MN-MATE platform. Proceedings of the 2012 International Workshop on Programming Models and Applications for Multicores and Manycores[C]. 2012.
- [60] Ramos, Luiz E, Eugene Gorbatov, Ricardo Bianchini. Page placement in hybrid memory systems. Proceedings of the 2011 international conference on Supercomputing[C]. 2011.
- [61] Shin, Dong- Jae, et al. Adaptive page grouping for energy efficiency in hybrid PRAM- DRAM main memory. Proceedings of the 2012 ACM Research in Applied Computation Symposium[C]. 2012.
- [62] Zhaoguo Wang, Hao Qian, Haibo Chen, Jinyang Li. Opportunities and pitfalls of multi-core scaling using Hardware Transaction Memory. Proceedings of Asia-Pacific Workshop on Systems (APsys 2013) [C]. Singapore, 2013.
- [63] Qureshi, Moinuddin K, et al. PreSET: improving performance of phase change memories by exploiting asymmetry in write times[J]. ACM SIGARCH Computer Architecture News. ACM, 2012(40): No. 3.
- [64] Justin Meza, et al. A Case for Efficient Hardware/Software Cooperative Management of Storage and Memory. Proceedings of the 5th workshop on energy efficient design[C]. 2013.
- [65] M K Qureshi, V Srinivasan, J A Rivers. Scalable high performance main memory system using phase-change memory technology. In ISCA'09[C]. 2009.
- [66] Ran Liu, Heng Zhang, Haibo Chen. Scalable Read-mostly Synchronization Using Passive Reader-Writer Locks. Proceedings of Usenix Annual Technical Conference (Usenix ATC 2014)[C]. Philadelphia, USA, 2014.
- [67] Jie Fan, Song Jiang, Jiwu Shu, Youhui Zhang, Weimin Zhen. Aegis partitioning datablock for efficient recovery of stuck-at-faults in phase change memory. In MICRO'13[C]. 2013.
- [68] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. OSDI[C]. 2004.
- [69] ZooKeeper`http://zookeeper.apache.org/`.
- [70] Vinod Kumar Vavilapalli, Arun C Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, Eric Baldeschwieler. Apache Hadoop YARN: Yet Another Resource Negotiator. SOCC[C]. 2013.
- [71] Mike Burrows. The Chubby lock service for loosely-coupled distributed systems. OSDI[C]. 2006.
- [72] Sanjay Ghemawat, Howard Gobioff, Shun-Tak Leung. The Google File System. SOSP[C]. 2003.
- [73] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C sieh, Deborah A Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber. Bigtable: A Distributed Storage System for Structured Data. OSDI[C]. 2006
- [74] Daniel Peng, Frank Dabek. Large-scale Incremental Processing Using Distributed Transactions and Notifications. OSDI[C]. 2010.
- [75] James C Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, JJ Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, Dale Woodford. Spanner: Google's Globally-Distributed Database. OSDI[C]. 2012.
- [76] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman,

- Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall and Werner Vogels. Dynamo: Amazon's Highly Available Key-value Store. SOSP[C]. 2007.
- [77] HBase <http://hbase.apache.org/>.
- [78] Cassandra <http://cassandra.apache.org/>.
- [79] Nancy Lynch, Seth Gilbert. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services[J]. ACM SIGACT News, 2002, 33(2) : 51-59.
- [80] Michael Isard, Mihai Budiu, Yuan Yu, Andrew Birrell, Dennis Fetterly. Dryad: Distributed Data-Parallel Programs from Sequential Building Blocks. EuroSys[C]. 2007.
- [81] Yuan Yu, Michael Isard, Dennis Fetterly, Mihai Budiu, Ulfar Erlingsson, Pradeep Kumar Gunda, Jon Currey. DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language. OSDI[C]. 2008.
- [82] Hadoop <http://hadoop.apache.org/>.
- [83] Hadoop Streaming <http://hadoop.apache.org/docs/r1.2.1/streaming.html>.
- [84] Hive <https://hive.apache.org/>.
- [85] Mahout <https://mahout.apache.org/>.
- [86] S4 <http://incubator.apache.org/s4/>.
- [87] Storm <https://storm.incubator.apache.org/>.
- [88] Shivaram Venkataraman1 Erik Bodzsar2 Indrajit Roy Alvin AuYoung Robert S Schreiber. Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices. EuroSys[C]. 2013.
- [89] Spark <http://spark.apache.org/>.
- [90] Spark Streaming <http://spark.apache.org/streaming/>.
- [91] Spark SQL <http://spark.apache.org/sql/>.
- [92] GraphX <http://spark.apache.org/graphx/>.
- [93] MLlib <http://spark.apache.org/mllib/>.
- [94] MESOS <https://mesos.apache.org/>.
- [95] 廖小飞, 范学鹏, 徐飞, 李鹤, 金海. 数据密集型大规模计算系统[J]. 中国计算机协会通讯, 2011, 7(7).
- [96] 肖依, 赖明澈. 数据密集型计算系统结构[J]. 中国计算机协会通讯, 2011, 7(7).
- [97] 过敏意, 须成忠. 云计算的编程模型与自适应资源管理[J]. 中国计算机协会通讯, 2011, 7(7).
- [98] 宫学庆, 金澈清, 王晓玲, 张蓉, 周傲英. 数据密集型科学与工程: 需求和挑战[J]. 计算机学报. 2012, 35(8).
- [99] ADFS <https://github.com/taobao/ADFS>.
- [100] OceanBase <https://github.io/oceanbase>.
- [101] Higo <https://github.com/muyannian/higo>.
- [102] Zhou Wei, Guillaume Pierre, Chi-Hung Chi. CloudTPS: Scalable Transactions for Web Applications in the Cloud[J]. TPDS, 2011.
- [103] Jinquan Dai, Jie Huang, Shengsheng Huang, Bo Huang, Yan Liu. HiTune: Dataflow-Based Performance Analysis for Big Data Cloud. USENIX ATC[C]. 2011.
- [104] Rong Chen and Haibo Chen. Tiled MapReduce: Efficient and Flexible MapReduce Processing on Multicore with Tiling [J]. ACM Transactions on Architecture and Code Optimization (TACO). 2013, 10(1) : No. 3.

- [105] ICTBench <http://prof. ict. ac. cn/ICTBench/>.
- [106] Zhen Jia, Lei Wang, Jianfeng Zhan, Lixin Zhang, Chunjie Luo. Characterizing data analysis workloads in data centers. IEEE International Symposium on Workload Characterization (IISWC) [C]. 2013.
- [107] Lei Wang, Jianfeng Zhan, Chunjie Luo, Yuqing Zhu, Qiang Yang, Yongqiang He, Wanling Gao, Zhen Jia, Yingjie Shi, Shujie Zhang, Cheng Zhen, Gang Lu, Kent Zhan, Xiaona Li, Bizhu Qiu. BigDataBench: a Big Data Benchmark Suite from Internet Services. The 20th IEEE International Symposium On High Performance Computer Architecture (HPCA-2014) [C]. Orlando, USA, 2014.
- [108] Zhengping Qian, Xiuwei Chen, Nanxi Kang, Mingcheng Chen, Yuan Yu, Thomas Moscibroda, Zheng Zhang. MadLINQ: Large-Scale Distributed Matrix Computation for the Cloud. EuroSys [C]. 2012.
- [109] Zhengping Qian, Yong He, Chunzhi Su, Zhuojie Wu, Hongyu Zhu, Taizhi Zhang, Lidong Zhou, Yuan Yu, Zheng Zhang. TimeStream: Reliable Stream Computation in the Cloud. EuroSys [C]. 2013.
- [110] Toreca <http://djt. qq. com/articleview329>.
- [111] 王鹏, 孟丹, 詹剑峰, 涂碧波. 数据密集型计算编程模型研究进展[J]. 计算机研究与发展, 47(11), 2010.
- [112] Liang Yuan, Chen Ding, Daniel Tefankovic, Yunquan Zhang. Modeling the Locality in Graph Traversals. ICPP 2012 [C]. 2012: 138-147.
- [113] Yuxin Tang, Yunquan Zhang, Hu Chen. A parallel shortest path algorithm based on graph-partitioning and iterative correcting[J]. Comput. Syst. Eng., 2009, 24(157).
- [114] Neumeyer L, Robbins B, Nair A, et al. S4: Distributed stream computing platform. Proceedings of the 2010 IEEE International Conference on Data Mining Workshops [C]. 2010.
- [115] Rui Chu, Nong Xiao, Yongzhen Zhuang, Yunhao Liu, Xicheng Lu. A Distributed Paging RAM Grid System for Wide-area Memory Sharing, 20th International Parallel and Distributed Processing Symposium (IEEE IPDPS) [C]. Greece, 2006.
- [116] Redis, Mar. 2013. <http://www.redis.io/>.
- [117] Zhou, Ping, et al. A durable and energy efficient main memory using phase change memory technology [J]. ACM SIGARCH Computer Architecture News. ACM, 2009, 37: No. 3.
- [118] Ham T J, Chelepalli, B K, Xue N, et al. Disintegrated control for energy-efficient and heterogeneous memory systems. Proceedings of the 19th International Symposium on High Performance Computer Architecture [C]. 2013.
- [119] Jung, Ju-Young, Cho, Sangyeun. Memorage: Emerging Persistent RAM Based Malleable Main Memory and Storage Architecture. Proceedings of the 27th International ACM Conference on International Conference on Supercomputing [C]. 2013.

作者简介

陈海波 上海交通大学软件学院教授、博导, 主要研究方向为系统软件、系统结构与系统安全等。中国计算机学会高级会员。



廖小飞 华中科技大学计算机学院教授、博导，主要研究方向为系统软件、多核体系结构等。中国计算机学会高级会员。



罗英伟 北京大学信息科学技术学院教授，博导。中国计算机学会杰出会员，系统软件专委委员。主要从事计算系统虚拟化、云计算、地理信息系统等方面的研究。



软件定义的云数据中心网络研究进展

CCF 开放系统专业委员会

李 丹¹ 刘方明² 郭得科³ 何 源⁴ 陈贵海⁵

¹清华大学计算机系，北京

²华中科技大学，武汉

³国防科学技术大学信息系统与管理学院，长沙

⁴清华大学软件学院，北京

⁵南京大学计算机科学与技术系，南京

摘 要

软件定义网络技术是计算机网络领域近年来新兴的热门技术，而数据中心网络被认为是软件定义网络最重要的应用场景。本文综述了软件定义的云数据中心网络技术在学术界、工业界和标准化领域的国内外研究进展，并对其发展前景进行了展望。

关键词：软件定义网络，数据中心网络，云计算，虚拟网络

Abstract

Software Defined Networking (SDN) is one of the most important emerging technology in computer networks, and data center network is regarded as the most representative paradigm for applying SDN. This paper surveys the recent progress about SDN technology in academia, industry and standardization, and discusses the development prospective.

Keywords: software defined network, data center network, cloud computing, virtualized network

1 引言

现代社会信息量的爆炸式增长、资源复用技术的成熟和宽带网络的普及，共同促进了云计算的诞生和发展。以亚马逊公司的 EC2、谷歌公司的 AppEngine 和微软公司的 Windows Azure 等为代表的云计算服务已经得到初步商用，使得云计算逐渐成为人们按需使用软硬件资源和进行大数据深度挖掘处理的新型计算模式。据工业和信息化部电信研究院 2014 年发布的《云计算白皮书》显示，2013 年全球云计算服务规模约为 1 317 亿美元，年增长率为 18%，据预测未来几年云服务市场仍将保持 15% 以上的增长率，2017 年将达到 2 442 亿美元。我国当前还处于云计算发展初期，虽然我国云计算服务的总体规模目前仅占全球市场的 4% 左右，但近几年一直呈上升趋势。随着大数据时代的来临，作

为大数据处理的重要技术手段，云计算的发展空间将更加广阔。Gartner 公司列出的“2014 年十大科技趋势”中，与云计算有关的技术就占了其中 4 项，分别是“软件定义网络（SDN）”、“更大的数据和存储”、“混合云”以及“向虚拟数据中心的演变”。

数据中心是云计算的核心基础设施。谷歌公司早在 2006 年底就在全世界建造了能容纳超过 46 万台服务器的分布式数据中心。Facebook 公司于 2011 年 4 月对外展示了其建造在俄勒冈的数据中心，其中拥有数以万计的服务器，并在节能减排方面进行了示范。我国各地方政府也积极建造数据中心以促成云计算产业落地，如呼和浩特云计算产业基地、南京软件开发云平台、镇江“云神”工程、无锡“云谷”等。中国电信、中国移动、中国联通等电信运营商都在积极利用其 IDC 数据中心打造云计算战略。百度、阿里巴巴、腾讯等为代表的互联网企业也在大力发展云数据中心，以提供更好的云计算服务。

数据中心网络在云计算基础设施中具有关键地位。云数据中心网络是连接数据中心大规模服务器的桥梁，也是承载网络化计算和网络化存储的基础。云计算的核心价值在于大数据的集中处理和资源的统计复用。由于大规模云计算任务往往伴随着服务器之间的海量数据交互，数据中心网络性能的高低决定了云计算的服务质量。同时，由于网络资源天然存在的共享特性，如何让云计算用户实现安全而公平的网络访问，也是关系到云计算用户体验的重要因素。当前的主流云计算提供商都积极探索面向云数据中心网络的创新技术以提高云计算的服务质量。例如，谷歌公司于 2012 年 4 月对外公布其数据中心采用 SDN/OpenFlow 技术进行网络管理。

近年来云数据中心网络不但是互联网服务公司和网络运营商重点关注的领域，而且成为各国政府、学术界、标准化组织和设备商共同关注的焦点。欧盟 FP7 计划资助了 TCLOUDS、A4CLOUD 等云计算和数据中心项目，美国 NSF 资助了“Storage Class Memory Architecture for Energy Efficient Data Centers” 和“Scheduling Energy Consumption in Green Datacenters” 等数据中心节能项目，日本总务省的 SCOPE 计划资助了 DependableCloud 等相关项目。最近几年的 SIGCOMM、OSDI、SOSP、ISCA 等系统与网络领域的顶级国际学术会议上都有大量的数据中心网络论文发表。其中，作为计算机网络领域的顶级国际学术会议，在 SIGCOMM 2011 和 SIGCOMM 2012 上，数据中心网络方向的论文数量更是达到论文总数的 1/3 以上。国际互联网标准化组织（IETF）成立了以云数据中心网络为主要应用场景的工作组 SDN（Software Driven Networking），IEEE 也成立了针对数据中心网络的任务组 DCB（Data Center Bridge）。思科、瞻博、华为等网络设备厂商先后推出了数据中心交换机产品。2012 年 7 月，VMware 公司收购软件定义网络（SDN）领域的先驱和网络虚拟化市场的领导者 Nicira 公司，进军云数据中心网络虚拟化市场。

软件定义网络（Software Defined Networking，SDN）是近年来涌现的新兴网络技术。SDN 的核心思想主要有两点：第一，提高硬件平台的可编程性，从而快速实现新型网络功能的配置，满足灵活多变的应用需求；第二，把网络控制层面与数据转发层面分离，把软件控制功能放到网络管理器上，从而提高网络的管理控制能力。对于 SDN 技术是否可应用于广域网环境，目前学术界和工业界还存在不少争议，但一般认为云数据中心网

络是 SDN 技术的理想应用环境。创建软件定义的可定制云数据中心网络基础设施，是提高网络性能、实现多租户网络共享以及控制网络能耗的基本保障。美国斯坦福大学提出的 OpenFlow 协议是当前最具代表性的 SDN 协议。然而由于 OpenFlow 协议存在数据转发流程过于复杂、转发设备处理功能非常有限等问题，当前学术界正在积极探索其他可能的 SDN 架构。如何设计支持可软件编程网络节点和可扩展控制器的软件定义网络框架，并基于此框架实现云数据中心网络的大规模横向扩展，是未来需要首先解决的重要挑战。

本文将主要叙述软件定义的云数据中心网络近年来的国内外研究进展，并展望其发展趋势。

2 国际研究现状

2.1 软件定义网络体系结构

2000 年以来，网络界对于大规模的网络试验的兴趣越来越高。受 Planetlab 和 Emulab 的启发，美国国家自然科学基金会（National Science Foundation, NSF）推出了大型试验网络 GINI^[159]，欧洲则提出了 EU FIRE program^[152]。在这种背景下，斯坦福大学于 2006 年启动了 Clean Slate（Clean-Slate Design for the Internet）项目，把研究重点放在 campus network 范围上。2007 年，斯坦福大学提出了 Ethane 项目，设计了一个逻辑上的集中式控制器，控制流的传输，从而实现了企业网络的权限控制功能^[154]。受 Ethane 的启发，Nick McKeown 等人提出了 OpenFlow 的概念，并在 2008 年 SIGCOMM 上发表了论文^[153] OpenFlow: Enabling Innovation In campus Networks。OpenFlow 的诞生意味着 SDN 雏形的形成。2009 年，MIT Technology Review 将 SDN 评为十大前沿技术^[155]，并且第一次使用了 SDN 这一名称^[152]。2011 年，专注于 SDN 发展的 Open Network Foundation（ONF）成立，它由谷歌、Facebook、微软、华为、德国电信等 155 家互联网公司、设备制造商、电信运营商、虚拟化厂商、测试厂商组成^[156]，标志着 SDN 逐步向着商用普及的方向发展。

SDN 被认为是由斯坦福大学 OpenFlow 技术逐步发展起来的，ONF 也采用 OpenFlow 作为 SDN 设计的重要组成部分。根据 ONF 提出的 SDN 三层体系结构，网络分为应用层（Application Layer）、控制层（Control Layer）和基础构造层（Infrastructure Layer），如图 1 所示。控制层向上提供北向接口 API，为应用提供网络服务。起初 ONF 并不希望为北向接口设立明确的标准以防阻止创新，但在 2014 年初，还是成立了 Northbound Interface Working Group，希望能够规范厂商混乱的北向接口定义。目前该 Working Group 还没有相关工作公布。相比较于北向接口，南向接口已经有比较成熟的方案，即 OpenFlow。OpenFlow 把分离开的 control plane 和 data plane 连接起来，实现集中式控制。

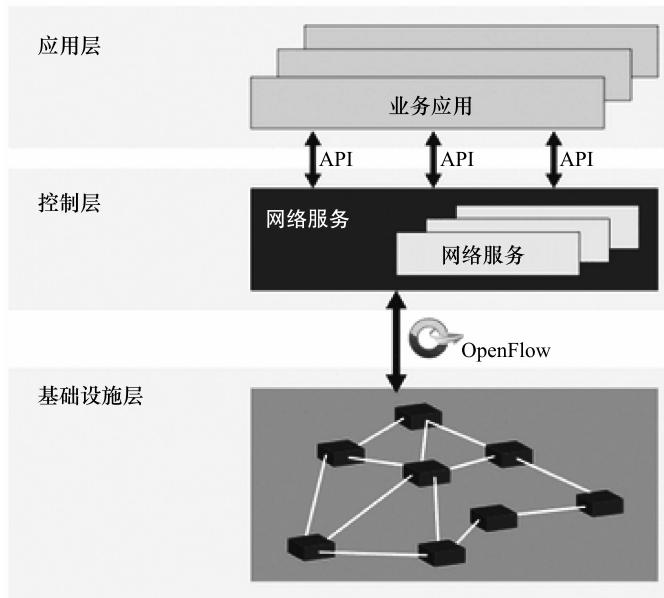


图 1 基于 OpenFlow 的典型 SDN 三层体系结构框架

SDN 借鉴了 active network 的网络编程思想。相比于 SDN 更强调网络控制器的可编程功能，active network 更多的是针对数据平面的可编程功能设计^[160,161]。SIGCOMM2013 年的文章^[162]对数据平面的可编程性进行了探讨。除了可编程性，针对 SDN 的数据平面的改进还有将计数器 counter 从 TCAM 中移出^[168]，用 CPU 作为数据平面的协处理器提高交换机性能^[169]，把部分常用转发信息放置于数据平面上，减少处理器负担^[170]等。

由于网络管理功能是一个全局性的工作，在 SDN 被提出之前，逻辑集中式的控制器^[164~166]已经进入研究人员的视野。服务器技术的发展使得用一个商用服务器即可存储一个运营商所有的路由状态并计算出相应的路由决策^[167]。但是，随着网络规模的扩大，集中式的控制器可能会带来控制消息延迟过大的问题，因此分布式的控制平面也成为重要的研究方向^[171~173]。对于控制平面来说，底层网络资源抽象化，控制平面可以像一个操作系统一样工作，网络操作系统（network OS）的概念应运而生，NOX^[173]就是其中的代表之作。

作为 SDN 的主要服务对象，网络应用才是最能体现 SDN 这一全新网络架构的价值之处。^[175,176]将 SDN 应用于 Middlebox 应用场景中。谷歌则将 SDN 应用到自己的数据中心网络架构中去^[177]。

现在 ONF 主导的以 OpenFlow 为标准的 SDN 设计可以视为狭义的 SDN 概念。广义的 SDN 概念认为 SDN 应该具备以下几个特征：分离的转发和控制平面、向应用开放的软件编程接口和集中式控制^[158]。其中最重要的就是可软件编程的开放接口，通过这些接口，应用可以与物理网络更贴合，更灵活地针对网络环境进行应变。而分离转发、控制平面和集中式控制则是为了实现这一功能而做出的设计^[158]。

云数据中心网络被认为是 SDN 的重要应用场景。下面的 2.2 ~ 2.6 节将对云数据中心

网络的国际研究进展进行综述。为了向读者展示相关工作的全貌，综述的内容中既包括与 SDN 有关的部分，也包括传统数据中心网络的部分。

2.2 云数据中心网络互联拓扑

云计算的规模需求决定了必须对数量众多的数据中心服务器进行互联，因此近年来学术界对云数据中心网络拓扑架构展开了广泛的研究。传统云数据中心普遍采用树型拓扑方案进行服务器互联^[60]。实践证明，这种拓扑方案已经不能很好地适应当前云计算数据中心的业务需求^[61~63]。

根据构建规则和互联技术的不同，当前的新型云数据中心网络拓扑可划分为五大类，分别是以交换机为核心的拓扑结构、以服务器为核心的拓扑结构、模块化数据中心、随机型数据中心和无线数据中心。如图 2 所示。

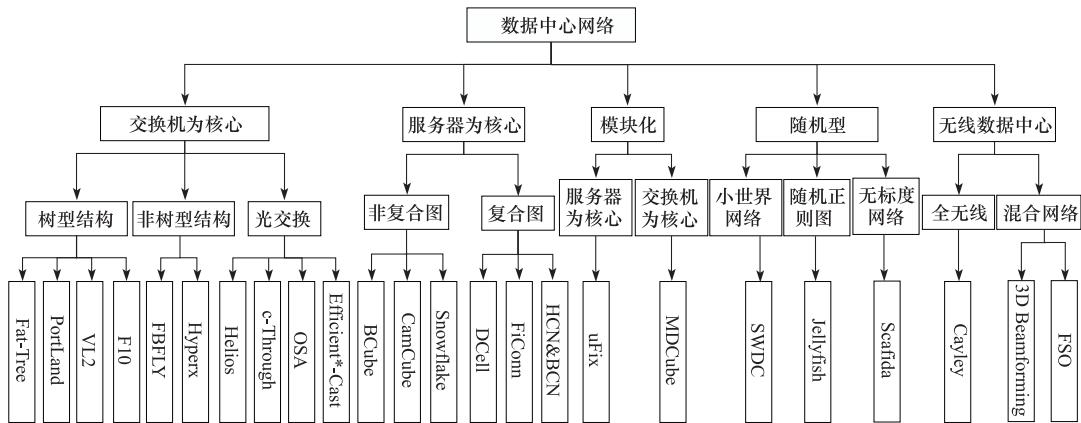


图 2 数据中心网络结构的分类系统图

(1) 以交换机为核心的云数据中心网络架构

在以交换机为核心的拓扑中，网络连接和路由功能主要由交换机完成。以交换机为核心的数据中心结构又分为三种类型，分别是树型结构、非树型结构和光交换数据中心结构。

在传统的树型结构中，交换机被划分为三层（接入层、汇聚层以及核心层），层间交换机两两相连而构成完全图，其弊端是存在顶层带宽瓶颈和单点失效。为此，Fat-Tree^[66]将接入层和汇聚层交换机划分为不同集群，保证无阻塞传输的同时消除单点失效问题。为了拓展虚拟化技术在数据中心中的运用以及增加网络灵活性，PortLand^[67] 和 VL2^[68] 分别修改交换机和服务器端协议以支持虚拟机迁移。同时，国内外学者在 Fat-Tree 基础上进行了再设计，王聪等人额外增加汇聚层和核心层以解决带宽瓶颈问题^[69]，F10^[70] 则打破 Fat-Tree 对称性以增强网络容错能力。

但树型结构面临扩展性不足、扩展成本高、容错性不强等挑战，严重影响系统性能。为此，研究者们另辟蹊径，摒弃传统的三层树型结构，基于大量同构的高密度端口商用

交互机构造扁平化结构。FBFLY^[71]和HyperX^[72]是其中的典型代表，它们都采用通用超级立方体结构互联普通的同构架构，实现网络拓扑结构的可扩展、高带宽以及高容错。

光交换技术以其高速率、稳定性、安全性等独特优势在数据中心中得以运用。Helios^[73]和c-Trough^[74]分别在树型结构的核心层和接入层中引入光交换机，使网络中并存光纤链路和分组链路。OSA^[75]则摒弃电信号交换机和分组链路，引入光交换矩阵和波长选择开关，电信号在机架顶端转换为光信号，到达目的地之后又转换为电信号。然而前三者只把光器件作为加速设备，Efficient * - Cast^[76]真正让光器件参与到结构构建之中，通过对光器件的动态重组来支持不同模式流量。

(2) 以服务器为核心的云数据中心网络架构

以服务器为核心的拓扑结构主要依靠服务器实现互联和路由，服务器通过多个网络接口接入网络。以服务为核心的拓扑结构通常采用递归方式构造，高层网络由多个低层网络互联而成，实现逐级扩展。我们将该类型结构分为复合图^[77]（Compound Graph）结构和非复合图结构以揭示其拓扑特性。

1) 基于复合图的网络结构：代表性工作有Dcell^[63]、FiConn^[78]、HCN以及BCN^[77]。Dcell的构建过程中要求低层的每台服务器分别与其他低层网络中的相应服务器相连，更重要的是，低层网络的个数必须等于低层网络所包含的服务器个数加一。若把每个低层网络结构看作一个节点，则高层网络结构是这些节点的完全图。DCell宏观上采用了层次化的完全复合图指导拓扑结构的设计，而且是递归定义的拓扑结构。其优点是可扩展性好，假如使用6口的小型交换机构建三层DCell网络，最多可以互联3263442台服务器。

2) 基于非复合图的网络结构：除复合图外，研究者还采用其他设计方法设计以服务器为核心的数据中心网络。Bcube^[79]采用递归构建，形成通用超级立方体结构，每增加一层，网络容量呈指数增长；Camcube^[80]在Trous结构的基础上，着重解决外部服务定制协议的实现问题；雪花结构^[81]采用科赫曲线，以服务器为核心递归扩展。

(3) 模块化云数据中心的网络拓扑结构

模块化数据中心将小规模服务器集成到拥有冷却、维护、传输等功能的集装箱容器当中，再将这些模块互联为大型数据中心。传统数据中心中，每个机架容纳20~40台服务器，机架顶端配置交换机，以此组成数据中心扩展的基本单元。随着数据中心的进一步发展，模块化数据中心已经替代机架成为构建大型数据中心的基本单元，模块化的优点是配置时间短、移动性能好、系统和能源密度更高、冷却和配置成本更低^[98]，是构建高效、可控、可管、即插即用、弹性数据中心解决方案，但随之而来的问题是如何将集装箱规模数据中心互联成大型数据中心网络。

MDCube^[82]用于连接采用BCube构建的同构数据中心集装箱。uFix^[83]也是为了对数据中心集装箱进行互联而提出的拓扑结构，但对集装箱内部所采用的拓扑结构没有限制，即可以连接异构的数据中心集装箱。

(4) 随机型云数据中心的网络拓扑结构

随机型云数据中心网络引入少量的随机连接将相隔较远的服务器或交换机互连，从而减小网络直径。上述规则性网络拓扑的互联规则要求过于严厉，难以同时保障数据中

心的渐进可扩展和容量平滑扩容。为此，学术界从网络科学中寻找突破口，设计新型的数据中心网络拓扑结构。研究者们基于小世界网络模型而设计了 SWDC^[84]，基于随机正则图设计了 Jellyfish^[88]，基于无标度网络模型设计了 Scafida^[89]。

(5) 无线数据中心的网络拓扑结构

此前的四类数据中心网络涉及大规模有线链路的正确铺设和接入，由此产生巨大的人力建设和维护成本。同时有线传输中带宽仍然是制约数据中心网络性能的重要因素。极高频（60GHz）无线通信技术和激光通信技术能提供高数据传输速率^[64,65]，将其运用于数据中心可大大降低布线成本，并能极大提高网络效能，进行构造无线数据中心网络。

若采用无线通信技术，便能轻松地解决布线成本问题，同时增强网络结构灵活性。当前提出的 Cayley^[90] 和 3D Beamforming^[91] 无线数据中心结构均采用 60GHz 无线通信技术。原因是 60GHz 拥有丰富的频谱资源（5 ~ 7GHz 未被使用）、极高速传输速率（Gbps 级）、强抗干扰性、高安全性、高度方向性等特性。只要处于主瓣覆盖区域内的接收器就能接收到信号，通过优化设置参数能提高数据中心的带宽和传输性能。Cayley 数据中心基于 Cayley 图构建，通过在服务器前后两端加入收发器直接实现互联；3D Beamforming 以机架为基本单元，通过增加反射面实现远距离单跳传输。FSO^[92] 是基于激光通信技术设计的数据中心结构，它按需动态地调整状态，使光束经过顶层平面反射实现单跳传输。

从表 1 中可以看出，比较流行的结构是多根树、超级立方体和复合图。最原始的 Folded-Clos 结构能提供充足的链路资源，但受制于单点失效和带宽瓶颈；多根树结构具有高对称性、构造简单等特点，但不足的是对顶层超额订购率；超级立方体顶点数呈指数增长，能为大型数据中心网络提供足够的网络容量；复合图能保留原有拓扑的局部和整体特性，递归定义使其网络规模达到指数增长。同时，部分拓扑采用 Trous 结构，Trouis 中节点与其相邻的节点都相连，如此规则的结构使每个节点拥有相同的度，不足是规模有限，适用于集装箱规模数据中心。数据中心网络拓扑研究成果非常丰富，各具优势，但也存在着诸多弊端。

表 1 当前数据中心网络结构拓扑属性一览表

结构名称	服务器/ 交换机为中心	结构属性	结构名称	服务器/ 交换机为中心	结构属性
Fat-Tree	交换机	Multi-Root Tree	Snowflake	服务器	Koch Curve
PortLand	交换机	Multi-Root Tree	BCube	服务器	Generalized Hypercube
VI2	交换机	Folded-Clos	MDCube	服务器	Generalized Hypercube
F10	交换机	Multi-Root Tree	CamCube	服务器	3D-Torus
FBFLY	交换机	Generalized Hypercube	SMDC	服务器	Small World Network
HyperX	交换机	Generalized Hypercube	Jellyfish	交换机	Random Regular Graph
Helios	交换机	Multi-Root Tree	Scafida	交换机	Scale-Free Network
c-Trough	交换机	Multi-Root Tree	Cayley	服务器	Cayley Graph
OSA	交换机	—	FiConn	服务器	Incomplete Compound Graph
Efficient * - Cast	交换机	Dynamic Architecture	3DBeamforming	交换机	—
Dcell	服务器	Complete Compound Graph	uFix	服务器	Incomplete Compound Graph
HCN	服务器	Incomplete Compound Graph	FSO	交换机	Dynamic Architecture
BCN	服务器	Incomplete Compound Graph			

2.3 云数据中心网络路由和传输协议

由于云数据中心网络存在链路资源密集、端到端带宽极高、端到端时延极低、流量不可预测等明显不同于广域网的特征，适用于广域网的许多传统互联网路由和传输协议在云数据中心网络中运行效率较低。因此，近年来学术界开展了对云数据中心网络新型路由和传输协议的研究。

在路由协议方面，当前的研究分别对云数据中心中单播和组播路由进行了深入的研究。对于单播路由，目前的研究成果主要可以分为两大类：基于流级别的路由方案和基于报文级别的路由方案。以 ECMP 为代表的传统的基于流级别的路由方案无法均匀地分配流量，导致网络出现热点。Hedera^[51]利用边缘层交换机对网络中的流量进行实时监测，一旦发现某条流的吞吐量超过预先设置的阈值，即将该条流标记为大流。一个集中式的控制器会周期性地从边缘层交换机获取大流信息，然后为所有大流计算路径并配置相应的交换机的路由表项。Mahout^[52]采用与 Hedera 类似的思想，差别在于 Mahout 使用每个连接对应的套接字缓冲区的使用情况来判断对应的流是否为大流。这样可以大大减小交换机表项的数量。Hedera 和 Mahout 都使用集中控制的方式对大流的流量进行周期性地调整，避免网络热点的出现。LocalFlow^[53]则是由每台交换机利用本地信息对经过的所有流进行分析。根据各个流的目的地和流量大小，将某些流拆分为若干子流、每条子流使用不同的路径进行传输，达到更高效的带宽利用率。

与流级别的路由方案相比，报文级别的路由方案能够实现更均匀的流量分配，更有效地利用链路资源，提高网络吞吐量。RPS^[53]是第一个报文级别的路由方案。RPS 利用基于 Clos 的数据中心网络的拓扑特点，提出将同一条流的报文以随机的方式分配到不同的路径上。由于不同路径的长度相等，因此各个报文经历的传输时延相差不会太大，在接收端不会产生严重的乱序现象。DRB^[54]则是利用基于 Clos 的数据中心网络的路由特点进行报文级别的路由。在基于 Clos 网络的数据中心网络中，只要路径中的层次最高的交换机被选定，则该条路径也被确定。因此 DRB 在发送端使用 IP 封装机制指定每个报文的传输路径，即将报文传输至层次最高的交换机。该交换机负责报文解封装，并将原始报文发送至目的端。

对于组播路由，由于云数据中心网络链路资源非常丰富，传统的接收者驱动的组播路由协议在云数据中心网络中构造组播树时，会浪费大量链路。ESM^[18]提出了一种新型的发送者驱动的组播路由协议，能显著提高组播树的链路利用率。同时，由于云数据中心网络普遍采用的低端网络设备路由表项较少，而组播路由表项又难以聚合，因此可扩展组播路由问题在云数据中心网络中尤其突出。MBF^[19]对标准的布隆滤波器（Bloom Filter）进行有效改进，并通过改进后的布隆滤波器来压缩组播路由条目，极大地提高了云数据中心网络组播路由的可扩展性。MCMD^[50]针对大量使用组播地址、造成网络设备组播转发表超出硬件限制的问题，优化了组播地址的使用：根据硬件资源的限制、组播组的规模和设置的策略，把应用层组播地址翻译为网络层组播地址或一组单播地址，从

而大大减少网络设备中组播转发表的条目，提高组播协议的可扩展性。Datacast^[59]利用数据中心网络丰富的链路资源，为每一个组播组使用多棵组播树加速数据传输。同时利用网络节点缓存数据提高组播的可扩展性。而文献^[57]则通过对网络设备分组、为不同分组的网络设备赋予不同组播地址块解决组播的可扩展性问题。

在传输协议方面，目前学术界主要对 TCP 协议进行改进，以适应云数据中心网络的特征。为了解决云数据中心分布式计算中典型的“多对一”传输模式下的 TCP InCast 问题，文献^[20]通过把发送方的超时重传定时器减小到微秒级别来减轻 TCP InCast 对系统性能的影响，DCTCP^[21]和 ICTCP^[22]则通过改进 TCP 的拥塞控制算法来提高端到端吞吐率。LTTP^[55]通过结合卢比编码和 TFRC 有效避免 TCP InCast 的发生，提高了“多对一”传输模式下的吞吐率。为了在云数据中心网络的多路径环境下提高端到端网络性能，多径 TCP (MP-TCP) 方案^[23]也引起了极大的关注。MP-TCP 在同一对源端和目的端之间建立多条连接，源端将数据拆分成若干部分，使用不同的连接同时进行数据传输，从而增加单位时间的数据传输量。另外，文献^[56]提出了一套解决 TCP 在数据中心网络中性能低下的方案。该方案包含 CP (Cutting Payload) 和反馈机制 PACK (Precise ACK)。CP 将交换机准备丢弃的包的所有净荷删除，但将头部信息传送给接收端。接收端收到被 CP 处理后的报文便可准备获网络拥塞情况，并通过确认报文通告给发送端。发送端基于这些信息，及时、准备地进行拥塞处理。

2.4 云数据中心网络虚拟化

虚拟化是保障云计算安全和实现资源复用的重要技术，因此虚拟化技术对云数据中心尤其重要。传统的计算虚拟化和存储虚拟化技术发展比较成熟，但网络虚拟化技术的进展则相对滞后。云数据中心的网络虚拟化是近年来学术界的研究热点，尤其是云数据中心虚拟网络隔离技术和虚拟资源分配问题。

当用户需要租用多个虚拟机，虚拟机之间便通过网络进行互联和通信，并组成了虚拟数据中心网络。出于安全考虑，不同用户的虚拟数据中心网络需要进行隔离，属于不同虚拟网络的虚拟机在缺省配置下不允许互相通信。NetLord^[24]提出将用户虚拟机的以太网分组封装在第三层 IP 分组上的方法来实现数据中心网络的虚拟化及多用户支持。通过封装，用户虚拟机使用的 MAC 地址不会出现在转发表中，不但解决了不同用户的 MAC 地址空间重叠的问题，而且大大减小了交换机的转发表空间。

为了适应数据中心资源共享和服务器整合的需要，数据中心的虚拟机应该具备实时迁移能力。传统的基于二层网络的虚拟机迁移方案受二层网络规模的限制，难以扩展基于三层网络的移动 IP 机制实现开销过大，难以适应大规模数据中心较为频繁的虚拟机实时迁移。近年来，IETF 制定的大二层路由协议 TRILL^[25]以及虚拟机迁移的报文格式标准 VXLAN^[26]和 NVGRE^[27]，尝试从不同的角度解决这些问题。TRILL 通过使用 MAC-in-MAC 封装，减少了交换机的转发表容量，并使用链路状态协议实现了二层网络上的高效路由。在 TRILL 协议的实现中，入口边缘交换机会对收到的二层数据分组

进行 MAC-in-MAC 封装，并在内层以太网头部外面增加一个自定义的 TRILL 头部。在这个 TRILL 头部中，定义了数据分组的入口边缘交换机和出口边缘交换机，并包含了一个用于避免路由环路的字段。VXLAN 会对二层数据分组进行 MAC-in-UDP 封装，并定义了一个新的 VXLAN 头部，插入到二层数据分组头部与外层 UDP 头部之间。这个 VXLAN 头部包含了一个 24 比特的字段，用于区分不同的虚拟网络。NVGRE 则是使用 GRE 标准封装格式对二层数据分组进行封装，并在 GRE 头部包含一个 24 比特的虚拟网络标识。

由于不同的虚拟数据中心网络共享同一个物理数据中心，设计公平而高效率的资源共享尤其是带宽共享机制非常重要。传统网络由服务器通过基于 TCP 流的竞争方式来实现网络带宽共享，但由于用户可以通过增加 TCP 流的数量来获得更高的带宽，因此这种资源分配方式在云数据中心网络中并不公平。目前主要有两种思路来解决这一问题。一是基于资源竞争的方案。其基本思路是在虚拟机或用户级别实现公平的带宽竞争，防止应用程序通过增加流数目的方式骗取网络资源。典型的基于竞争的带宽共享机制包括 Seawall^[28]、Netshare^[29] 和 Faircloud^[30] 等工作。二是基于资源分配的方案。基于资源分配的带宽共享方案有两种。一种是确切地定义每个虚拟机或者每个用户对网络带宽的需求，直接给虚拟机分配足量的带宽，并且通过限速机制来确保每个虚拟机或者用户对带宽的利用不会超过分配的限额。典型的带宽分配机制包括 SecondNet^[31]、Oktopus^[32]、TIVC^[33]、Bazaar^[34]、Hadrian^[149] 等工作。另一种是确切地定义每个虚拟机或者每个用户对网络带宽的需求的下限，通过限速机制及发送方与接收方的通信保障每条流对网络带宽需求的下限，让虚拟机或者用户竞争剩余的网络带宽。典型的带宽分配机制包括 EyeQ^[150]、ElasticSwitch^[151]。总体而言，相对基于竞争的带宽共享机制而言，基于分配的方案可以提供真正的带宽“保障”，但用户可能无法充分利用所申请的网络带宽，从而造成云数据中心网络资源的浪费。

2.5 云数据中心网络能耗管理

大规模云数据中心在支持日益丰富的应用和众多租户资源需求的同时，也带来了巨额的能耗成本以及碳排放污染问题，引起了学术界和工业界的高度重视。美国 NSF 资助了“Storage Class Memory Architecture for Energy Efficient Data Centers” 和“Scheduling Energy Consumption in Green Datacenters” 等数据中心节能相关的项目。前者的研究目标是采用 DRAM、PRAM、Flash 等多种技术融合的方案，来构造高效节能的云数据中心大容量存储系统。后者的研究目标是在数据中心使用太阳能和风能等绿色能源以减少对传统能源的使用，通过对数据中心负载和绿色能源供应的建模分析，设计合理的负载调度机制和绿色能源使用策略，将绿色能源与传统能源进行有机配合使用。

目前国际上从云数据中心网络能耗监控、设备节能管理和多能源规划等方面的研究方兴未艾，下面分别介绍相关研究进展。

微软研究院在 2009 年的 Genome 项目^[35] 中，提出对数据中心采用无线传感器网络进

行监控的方案。研究人员在微软的一个大型数据中心里部署了约 700 个传感器，对数据中心内各位置的温度进行监控并实时发现热点，为数据中心内工作负载的合理分配调度提供依据。IBM 公司也开展了与微软公司类似的研究项目^[36]，研究人员提出了 MQTT-S + MPERIA 的无线传感器网络通信协议。美国劳伦斯国家实验室也对无线传感器网络应用于数据中心节能作了广泛研究和实验^[37,38]。

云数据中心存在大量的链路资源，而大部分时候网络负载远低于峰值负载，因此造成许多网络设备能耗的浪费。为了降低云数据中心网络能耗，一种简单的方式是当网络设备空闲时将其置于休眠模式^[39]。然而该方法容易造成网络报文的丢失，会对网络性能带来较大的负面影响。Gunaratne 等提出了自适应链路速率的能耗调整机制^[42]，能够在全双工传输的以太网中根据链路利用率自动切换链路速率，并进而降低设备能耗。为了解决速率抖动问题，Gunaratne 等进一步提出了使用链路利用率阈值策略和超时阈值策略进行链路速率的联合控制^[43]。Heller 等提出一种网络级的云数据中心网络能耗管理器机制“弹性树”(Elastic-Tree)^[44]，该机制能够根据云数据中心网络的负载情况，动态调整处于开启状态的网络链路和交换机，使用一个节能的网络拓扑子集来完成数据传输工作。

降低能耗并不等同于实现绿色计算，因为当前数据中心消耗的仍然主要是传统的高碳排放量的能源。绿色和平组织(GreenPeace)定义实现绿色信息技术的方式是“高能效加新能源”^[45]。因此，还需要充分利用新能源为云数据中心供能，例如 Green House Data 建在美国怀俄明州的风能供电数据中心和 Facebook 建在俄勒冈州的太阳能数据中心。eBay 公司使用 30 个 Bloom Energy 的燃料电池来为其在犹他州的数据中心供能。但为数据中心因地制宜制定多能源组合和配额规划方案面临着非常大的挑战。加州大学圣克鲁兹分校的 Jose Renau 等提出的 ReRack^[46] 可用来评估使用新能源的数据中心的能耗开销。宾夕法尼亚州立大学的 BhuvanUrgaonkar 等提出的碳感知能源规划方法可以帮助数据中心操作人员设计可持续发展的新能源驱动的系统^[47]。HP 实验室的 Daniel Gmach 等提出了一种能耗管理规划方案^[48]，使得数据中心的负载能耗需求与供应相匹配。Rafael Diaz 等提出了一种系统最佳能源组合模型^[49]，充分考虑了新能源对公众健康及地区经济效益的影响。

2.6 SDN 在工业界的发展现状

2.6.1 芯片厂商

(1) 盛科

2013 年 3 月，盛科正式对外发布新一代以太网交换芯片 GreatBelt (CTC5163) 系列，该芯片针对 SDN 做了多项创新。凭借该自主研发核心芯片的创新 N-FlowTM 技术，在 2013 年 4 月的美国 ONS 峰会上，第二代 OpenFlow 交换平台 V350 战胜 Big Switch、HP、NTT 以及 Redware 等强大对手中脱颖而出，夺得冠军，成为首届 SDN Idol；2014 年 5 月

27 日，V350 OpenFlow 交换平台顺利完成 ONF plugfest OpenFlow 1.3 测试。

盛科最新的 OpenFlow 交换平台 V350 基于盛科最新发布的第三代以太网芯片 CTC5163 和 N-Space 开放软件，采用创新的 N-FlowTM 技术，集成 OpenvSwitch 和盛科 SDK。它能提供高达 240Gbps 线速转发能力，整机系统功耗小于 60W，并且支持丰富的 OpenFlow V1.3 版本，芯片支持 OpenFlow 多级流表并将芯片内置的流表数扩大到相应产品的 16 倍，它有多达 64K 的精确匹配流表，通过将模糊匹配和精确匹配有机结合，充分发挥 OpenFlow 交换机的优势。作为支持 OpenFlow 的交换机参考设计，V350 提供了从核心芯片到系统软硬件的整体解决方案。依托 V350 良好的软硬件接口及平台开放性，用户可轻松打造定制的 SDN（Software Defined Network）解决方案，实现各种网络虚拟化应用。更为重要的是，V350 平台的开放性可以给 SDN 厂商提供差异化的定制方案，帮助其创新。

盛科作为的 SDN 芯片和白牌（Whitebox）设备提供商，还发布了业界首个基于硬件交换机的开源 SDN 项目——Lantern，Lantern 旨在为基于 SDN 硬件交换机提供 SDN 实现。Lantern 中集成了开源的 Linux Debian 7.2 OS、优化了的 Open vSwitch（OVS）、适配层以及芯片 SDK，Lantern 选择 GitHub 作为源码托管平台，使用开源 Apache 2.0 许可协议。

通过开放标准实现 SDN 极大地提高了灵活性，同时降低了研发成本。Lantern 提供包括芯片 SDK 和适配层在内的所有开源代码，为研究 OpenFlow 提供了极大的便利，同时设备厂商可在此基础上进行更多创新，此外，Lantern 也致力于为开放的生态系统做出贡献，这对整个 SDN 领域的发展至关重要。

（2）博通

全球有线及无线通讯半导体创新方案领导厂商博通（Broadcom）公司宣布推出 XLP® II 系列最新产品 XLP500 系列多核心通讯处理器。XLP500 系列搭载 32 NXCPU，并可达到 80Gbps 的效能，相较于竞争产品，每个核心提供最多至四倍的效能。

XLP500 系列提供卓越的处理效能与弹性，可简化网络功能虚拟化（Network Functions Virtualization，NFV）与软件定义网络（Software Defined Networking，SDN）的部署。此卓越效能归功于 4 指令执行（quad-issue）与 4 执行绪（quad-threaded）的超纯量架构与乱序执行功能。此产品也支持博通的 Open NFV 平台，并可与博通的 StrataXGS 交换器系列协同作业，因此能简化开发流程、达到最佳化功率、降低硬件成本，并缩短上市时间。

（3）英特尔

2013 年，英特尔围绕标准 x86 服务器和可编程交换单元构建了三个用于 SDN 和 NFV（网络功能虚拟化）开发与部署的参考设计，其中包括开放网络交换机平台、开放网络服务器平台，以及数据平台开发套件。

其中开放网络交换机平台基于可扩展的英特尔至强或英特尔酷睿处理器、英特尔 6700 系列以太网交换芯片，以及英特尔 89xx 系列通信芯片组构建。它实现了一个完整的基于英特尔架构的全新网络平台，支持 48 口的交换机，同时集成了计算能力、通信能力

以及 10G 和 40G 的网络交换能力；开放服务器平台参考设计基于英特尔至强处理器和英特尔 82599 万兆位以太网控制器，以及英特尔 89xx 系列通信芯片组，支持虚拟设备工作负载，运行于标准英特尔服务器。

2.6.2 设备和解决方案提供商

(1) 华为

2013 年 8 月，华为发布敏捷网络架构及全球首款敏捷交换机 S12700，旨在满足云计算、BYOD 移动办公、SDN（软件定义网络）、物联网、多业务以及大数据等新应用对更高可靠、大带宽、更大规模以太网的要求。该产品采用全可编程架构，能灵活快速满足客户定制需求，助力客户平滑演进至 SDN 网络。该产品基于华为公司首款以太网络处理器 ENP，内置随板 WLAN AC 无线局域网接入控制器，实现有线无线真正融合；它支持 iPCA 网络包守恒算法，对任意业务流随时随地逐点检测，助力客户对业务的精准管理。该产品基于华为公司自主研发的通用路由平台 VRP，在提供高性能的 L2/L3 层交换服务基础上，进一步融合了 MPLS VPN、硬件 IPv6、桌面云、视频会议等多种网络业务，提供不间断升级、不间断转发、CSS2 交换网硬件集群主控 1+N 备份、硬件 Eth-OAM/BFD、环网保护等多种高可靠技术，在提高用户生产效率的同时，保证了网络最大正常运行时间，从而降低了客户的总拥有成本 (TCO)。

另一方面，中国电信携手华为率先完成 IDC 网络部署 SDN 技术。双方合作于 2014 年 4 月在武汉成功完成中国首个传送网 SDN 解决方案测试，验证了现有存量网络向 SDN 演进的解决方案，标志着 SDN 的商用化进程取得了重要进展。

传送网 SDN 可实现网络集中化的管理，克服了传统网络分散控制的弊端，能帮助运营商快速建立灵活、开放和差异化的传送管道。该测试基于华为现有 OTN/ROADM 商用设备，采用集中式控制器，通过华为 T-SDN 方案成功实现了光层业务和电层业务的快速下发和在线网络评估，验证了华为 T-SDN 在业务运营自动化上给运营商带来的巨大提升。

(2) 思科

思科利用 Insieme 技术将一系列的交换设备与专有的“Insieme”专用集成电路 (Application-Specific Integrated Circuit, ASIC) 结合起来，用以改造网络“NX-OS”，又名基于 ACI 的策略控制器 (ACI-based policy controller)。专有 ASIC 可以让思科通过名为“原子计数器”(Atomic Counters) 的技术和敏捷的技术实现网络性能有保障的实时数据传输。专有 ASIC 对更多端点规模的应对能力优于现有的商业解决方案，并可同时为管理物理网络和虚拟网络提供一个统一的界面。ACI 可将基于软件的叠加层总拥有成本降低 25%，将能耗和冷却成本降低 15%。此外，ACI 还能够将应用部署时间由“数月缩短至数分钟”，同时为物理和虚拟化基础设施提供实时可见性。

(3) VMware

VMware 公司在 2013 年度 VMworld 大会上发布了其收购 Nicira 后推出的 SDN 核心产品 NSX。NSX 是一个独立的 hypervisor 云管理网络虚拟化平台，可以提供完整的 2~7 层

网络虚拟化服务。NSX 是 VMware 软件定义数据中心的一部分，是一个虚拟化的网络和安全软件产品，由 VMware vCloud 网络、安全（vCNS）和 Nicira 网络虚拟化平台（NVP）共同创建。NSX 提供了 VMware 虚拟化方面的云计算技术。

NSX 提供的虚拟化网络环境，不需要管理员对命令行接口直接干预，而是将网络操作从底层硬件抽象到一个分布式虚拟层，很像用于处理能力和运营系统的服务器虚拟化。VMware vCNS（以前称为 vShield）虚拟化网络的 4~7 层，而 Nicira 的 NVP 虚拟化 2~3 层。

（4）Juniper 瞻博

Juniper 瞻博公司在 2012 年 12 月收购了新兴企业 Contrail Systems，随后推出了商业版的 Contrail 和开源版的 OpenContrail 两款 SDN 解决方案。两款产品都是建立在相同的代码基础上，商业版本更注重用户在生产上的可靠性，并且 Juniper 会提供相应的服务。Contrail 将各类网络环境拆分为四个控制层——管理层、服务层、控制层以及转发层，将一部分功能集中到控制器之内，同时也为网络环境中的交换机及路由器提供一些其他功能。

OpenFlow 控制器与 Contrail 之间的最大区别在于，Contrail 会把转发列表的主副本保存在控制器当中，并将其复制到交换机端。相比之下，OpenFlow 控制器会将主副本保存在交换机当中，并在其内容发生变更后将其聚合到控制器内。

2.6.3 互联网公司

（1）Facebook

Facebook 具有自主设计的代号为“Wedge”的网络交换机，这款采用了模块化设计理念的新型网络产品，将可以非常方便地更换设备组件，来让设备运行自有的定制化软件，这与思科等网络提供商的传统网络交换机大不相同。

Facebook 已经开始使用 Wedge 网络交换机来为 10 亿多 Facebook 用户提供服务。对于日益扩张的 Facebook 内部运营所需的高效数据中心网络来说，Wedge 将会简化对其的创建和管理过程。

Wedge 围绕一个与核心服务器相同的微处理器模块搭建，在实际使用过程中，Facebook 不仅能为服务器和网络交换机配备同样的硬件，还可以在交换机上运行与服务器同样的操作系统。这套共用的操作系统也是 Facebook 自主研发的，可以大幅降低 Facebook 计算中心所需的人力资源。

（2）腾讯 SRP

为优化数据中心内部网络，腾讯通过构建新的 SRP 协议（Sequoia Routing Protocol，腾讯自研的路由协议），根据数据中心网络固定简单的 CLOS 架构组网和基于运营的预先规划的子网特点，通过预设好的静态路由以及根据邻居状态信息动态解析相结合的方式生成实际可用的路由转发表项。

腾讯和华为在 SDN 应用方面已展开合作，基于华为的 H3C SDN 解决方案，在 H3C SDN 交换机中运行腾讯自研的路由协议 SRP，使得网络路由收敛效率提升超过 50%，效

率显著提升，SRP 对网络规模增长不敏感，能够有效地支持数据中心规模继续扩大。

3 国内研究进展

本文很难全方位覆盖国内数据中心网络的研究状况，因此仅介绍一些具有代表性的研究成果，这些方案的技术思路见前面的章节。

(1) 微软亚洲研究院

由于认识到数据中心网络是数据中心基础结构中不可或缺的重要一环，微软亚洲研究院从 2007 年起开始进行数据中心网络方面的研究。研究主要涉及数据中心网络拓扑结构、传输协议、虚拟化等方面。

在拓扑结构研究中，微软亚洲研究院提出了以服务器为中心的网络结构 DCell^[63]、Ficonn^[78]、BCube^[79]、MDCube^[82]。在传输协议方面，提出了解决 TCP Incast 的方案 ICTCP^[84]。在虚拟化方面，发表了该方向的第一篇论文 SecondNet^[85]。

此外，微软亚洲研究院的研究人员还注意到，数据中心网络研究需要一个通用的网络平台，供大家实现各类创新设计。以软件进行分组转发的平台无法满足分组转发性能上的要求（带宽和时延），而当前的商用网络设备还无法进行编程。为了同时满足高性能和高可编程性，微软亚洲研究院设计并实现了 ServerSwitch^[86] 网络平台。ServerSwitch 综合了商用以太网交换芯片的高性能和通用 X86 芯片的可编程性，为未来的各类数据中心网络研究提供了一种新的平台。

(2) 清华大学

清华大学在数据中心网络拓扑结构、传输协议、无线通信、虚拟化、节能机制、增强以太网等方向均进行了相关研究。获得的相关项目资助包括国家 973 计划青年科学家专题项目“软件定义的云数据中心网络基础理论与关键技术研究”等。

在网络拓扑结构方面，清华大学提出了连接异构集装箱数据中心的网络拓扑 uFix^[83]，是第一个试图解决数据中心服务器和交换机异构性的拓扑方案。在传输协议方面，清华大学设计了采用 Bloom Filter 并结合数据中心特性进行可扩展组播路由的组播协议 ESM^[34]、MBF^[94]。通过建模揭示了 TCP Incast 现象的机理^[95]，还设计了通过编码传输来解决 TCP Incast 问题的方案^[96]。

在无线通信方面，清华大学研究了无线数据中心网络中的信道干扰问题^[97]，并且考虑了自适应速率^[98]进行全面建模与优化。在增强以太网方面，清华大学采用相平面^[99]和滑模结构^[100]分析了数据中心增强以太网中的流量控制机制。在节能机制方面，提出了数据中心网络节能路由协议 EAR^[101]，并比较了典型数据中心网络拓扑的能耗特性^[102]。

在软件定义网络方面，清华大学开发了一个系统 SODA^[103]，通过提供丰富的语义和更多灵活的可编程接口增强了 SDN 的数据平面。而且提出了软件定义的绿色数据中心^[104]，应用软件定义网络的技术，为每一条路由路径使用专有的路由提高了路由路径

的链路利用率。而且利用软件定义网络技术，提出了一种新型数据中心网络架构^[105]，通过给服务器和路由器都增加无线网卡来消除冗余流量。

(3) 华中科技大学

华中科技大学在云数据中心网络虚拟化与性能保障，以及节能减排与运营成本优化等方向开展了基础研究与系统实践。获得的相关项目资助包括与清华大学联合承担的国家973计划青年科学家专题项目“软件定义的云数据中心网络基础理论与关键技术研究”等。

在面向云计算性能保障的网络虚拟化方面，首次运用博奕论为IaaS云计算多租户环境下的数据中心网络带宽共享机制奠定了理论基础^[106]，并进一步结合软件定义的网络架构，实现了基于OpenFlow的虚拟机间分布式带宽分配协议Falloc^[107]，以及首个快速适应流量变化、抗短流干扰和短时网络拥塞的在线演化式带宽分配算法^[108]。该方案能够兼顾各租户基本带宽保证、按比例带宽分配公平性，以及云运营商整体系统利用率要求。为了综合解决计算、存储和网络等各个维度资源竞争造成的虚拟化性能开销，对IaaS云计算环境下大量虚拟机的性能管理研究给出了全面总结、方案对比及优化思想^[109]，并据此开发了性能干扰感知的虚拟机高效迁移机制iAware^[110]，在实现数据中心负载均衡及节能降耗等目标的同时，有效缓解对数据中心作业性能的影响。

在面向绿色计算的数据中心节能减排方面，建立了一套理论完备的数据中心“效能双赢”控制体系^[111]，从而能够量化数据中心能耗管理的全局设计空间和最优理论上限，进而灵活权衡数据中心的多种设计选择和当前云运营商迫切关注的单位能耗性价比指标。由于认识到单纯降低能耗并不等于实现绿色计算，发表了国内第一篇关于清洁可再生新能源在数据中心基础设施中应用的综述文献^[112]，并设计了数据中心多路能源供能系统的优化调度算法SmartDPSS^[113]，从而能够协同互补利用智能电网、间歇性不稳定新能源和有限容量不间断电源UPS。随着越来越多数据中心在全球范围的跨域分布式部署，提出了综合考虑能耗成本和碳排放量的地理分布式数据中心负载均衡算法^[114]以及应用新兴燃料电池的绿色数据中心多目标共赢调度机制^[115]。

(4) 中科院计算所

中科院计算所在数据中心网络关键设备、管理系统以及基于数据中心的未来网络体系架构设计方面开展了相关的研究工作。

在数据中心网络关键设备方面，设计并研制了可编程虚拟化路由器平台PEARL^[106]支持数据平面与控制平面灵活可编程，可以根据DCN网络不同拓扑、寻址和应用的需求，用软件方式定义数据包格式与转发方法。同时，可编程虚拟化路由器支持转发资源的虚拟化与隔离，可以有效保障不同类型业务的性能，提高网络转发设备的资源利用率。在此基础上，针对虚拟路由器上基于三元内容寻址存储器(TCAM)的IP查询机制和基于静态随机接入存储器(SRAM)的IP隧道查询机制会在路由更新时出现丢包的现象，根据转发信息(FIB)构建特里结构，叶子节点用TCAM进行查询，其余用SRAM进行查询^[117]。在网络体系结构方面，基于数据中心提出了面向服务的未来互联网体系架构SOFIA^[118]。SOFIA所包含的很多关键思想(如服务迁移和本地化、服务路由、网络休眠

等)都是数据中心网络相关技术的扩展和延伸。基于网络绩效受到缓存效率的影响这一观察,提出利用网络编码和随机转发策略来提高多径转发中的缓存效率,并以此设计了CodingCache^[119]。

(5) 西安交通大学

西安交通大学在软件定义网络数据通路、数据中心网络虚拟化、自动化管理配置、高效路由寻径等方面进行了相关的研究工作。

研究人员设计实现了从硬件到软件全可编程的软件定义网络交换机^[120],在灵活性和性能上取得较好的折中。基于提出的软件定义网络交换机构建的桌面式数据中心^[121],可提供灵活且低功耗的一体化数据中心实验平台。通过SDN的思想,提出基于带宽感知的多租户云数据中心虚拟网络分配算法^[122],实现网内带宽和网际带宽的保证,使租户网络应用的性能和服务弹性达到良好效果。利用现有数据中心网络的真实流量统计信息,对数据中心网络内流量拥塞及其发生的位置、持续时间、稳定性等进行了深入的分析,基于此提出了VirtualKnotter方法用于实时调整虚拟主机的位置,以对数据中心网络的流量拥塞进行控制^[123]。

针对大量存在的基于IP编址的数据中心网络,提出了DCZeroconf机制和协议,解决多个子网/虚拟网络并存的地址自动配置问题^[124];提出ETAC系统和相应算法将容错配置为难题建模到导出子图同构问题,可支持任意网络编址,提供了支持容错的自动数据中心网络管理配置方案^[125]。在路由寻径方面,通过对现有数据中心网络的连接方式和路由方式的深入总结和思考,指出了数据中心网络在寻找路径、流量工程、多播组播、路由安全方面存在的诸多问题和可行的研究方向^[126]。

(6) 国防科技大学

国防科大在数据中心网络可扩展拓扑、传输协议、节能机制等方向进行了卓有成效的研究。

在网络拓扑结构方面,提出了HCN、BCN、DCube三种具备无损可扩展和持续可扩展能力的以服务器为中心的互联结构^[77,127],并针对MapReduce的数据处理机制为数据中心设计了新型互联结构HFN^[128]。在传输协议方面,结合BCube等新型互联结构的拓扑特性研究了对Incast和Shuffle传输模式进行网内数据流量聚合的问题,提出了高效构造Incast和Shuffle数据流聚合结构的方法^[129,130]。在节能机制方面,提出了数据中心服务器层面的并发任务在线节能实时调度算法TL-DVFS^[131],除保证偶发任务集的最优可调度性外,还能达到实时约束与能耗节余之间的合理折中。在数据中心中控制器研究方面,提出了分布式控制器的最小覆盖方法、控制器的内容同步方法^[132]。

(7) 复旦大学

复旦大学在基于开放架构的高可靠软件定义网络体系研究、基于软件定义网络的关键设备研制及示范应用、基于软件定义网络的大规模分布式存储系统关键技术研制与示范应用等方向进行了相关研究。

在当前数据中心网络的典型结构中,添加了一个智能Middle Box,利用Clos网络构建SDN中间层,提升数据中心QoS和资源利用效率^[133];将复杂网络的思想用于数据中

心网络拓扑设计，Decluster 系统在可扩展性和性能方面取得更好的效果^[134]。在 SDN 应用研究方面，针对 Web 服务中的服务资源有限与终端差异等现状，分别从链路层和应用层两个层次，设计和实现了一种 Web 服务中的网络负载均衡方案^[135]。在节能机制方面，利用 SDN 的框架，提出包括流量优化单元和设备控制单元的节能模型，通过流量会聚、利用相对少的网络设备来满足通信需求^[136]。

(8) 北京邮电大学

北京邮电大学在数据中心网络带宽共享机制、TCP、多路径路由算法、多资源约束的虚拟机放置/迁移方法等方面展开了相关研究。

在数据中心网络带宽共享机制方面，提出了一种运营商友好的数据中心间数据块传输调度方法^[137]，并设计基于多属性信息的数据中心间数据传输调度方案。在 TCP 协议方面，通过减小 MTU，缓解 TCP Incast 问题引起的吞吐量下降并提高 TCP 流的公平性^[138]。在数据中心网络虚拟化方面，提出了一种基于内存压缩技术的虚拟机实时迁移方法^[139]。在虚拟机迁移方面，提出 SmartShuffle 来解决虚拟机在线迁移给数据中心带来的流量问题^[140]；在虚拟机的放置上，基于启发式算法对网络性能进行优化和降低物理机能耗，从而实现网络性能和能源之间的权衡^[141]；为了同时实现多种资源利用率优化，结合最小割原理提出一种贪婪算法，得到近似解^[142]。同时，为了降低运行成本，在不需要先验知识的条件下，提出采用 Lyapunov 优化框架来实现数据中心网络的最优控制，并对成本和延迟进行权衡^[143]。

(9) 大连理工大学

大连理工大学在数据中心的网络拓扑结构和网络交换机等方面开展研究。在数据中心网络拓扑结构方面，针对当前互连拓扑所存在的问题，设计了两种新型互连网络拓扑结构：交换交叉超立方体^[144]和交换折叠超立方体^[145]。其中交换交叉超立方体结构的网络直径大大缩减，具有硬件成本消耗低等优势；交换折叠超立方体结构具有更好的负载均衡特性。在数据中心网络交换机方面，提出基于负载均衡架构的 Byte- Focal 交换机^[146]。与现有的负载均衡交换机相比，Byte- Focal 交换机降低了实现复杂度，解决了数据包乱序问题并极大的减小了数据包延迟。同时，可以实现 100% 吞吐量。

(10) 东北大学

东北大学在数据中心网络结构等方面开展了相关研究。设计了数据中心网络结构^[147] Fat-Tree，网络由同构可编程交换机组成，中间的服务器将网络分成两个 Fat- Tree 变体。同样，每个 Fat-Tree 变体内部包含核心层、汇聚层和接入层，以保证每台服务器的任意网络端口都能同时以端口所允许的最大带宽进行通信，而不受网络通信带宽瓶颈的制约。该结构可容纳的服务器数量取决于交换机的端口配置，其优点是任意两台服务器之间的路径众多，提供了高连通性和吞吐量，但比 Fat- Tree 使用更多的交换机，硬件成本不容小觑。

(11) 开网科技

开网科技主要基于 SDN 技术开发计算机网络和网络安全教学与实训平台。通过 SDN 开放网络的特点，把网络内部行为通过完全可视化、可追踪的方式展现给用户，同时用

户可以自由地对网络功能进行配置和实验。采用基于 SDN 的网络虚拟化技术，每个用户可以独享一个虚拟网络，不同用户之间的虚拟网络互不干扰。开网科技公司的产品包括：基于 SDN 的软件化高速网络处理平台、高性能网络控制器及云网络管理平台、网络教学实训云平台、信息安全教学实训云平台。

(12) 南京叠锶

南京叠锶目前主要面向科研以及小规模试点部署提供开放网络设备，着重开展 SDN/OpenFlow 互连和应用方案。产品包括 x86 平台下的 PCIe 加速卡 ONetCard、ARM 平台下的 ONetSwitch 系列，前者着重于可编程逻辑（硬件）与高性能 CPU（软件）并举的单点性能，后者着重于数据平面的硬件加速以及成规模的网络互连与拓展。

4 国内外研究进展比较

当前国内关于数据中心网络的研究基本与国际学术界保持同步，在部分技术方向甚至处于领先地位。而且，国内云计算和大数据产业的迅猛发展，以及政府和企业对数据中心基础设施建设的大举投入，也为研发具有自主知识产权的数据中心网络核心技术提供了大规模平台和示范应用的潜在条件。

然而，国内工业界与学术界在产学研合作方面的紧密程度尚不如国外，导致目前国内高校和研究机构缺少在实际大型数据中心中实践创新方案并验证原型系统的机会，同时也利于企业及时把握前沿技术。因此，加强国内工业界与学术界在数据中心与 SDN 领域的深入交流与双赢合作，通过向学术界合理开放部分数据中心实验平台，从而吸引和推动先进理论与学术成果到实际系统的应用转化，将是我国在新兴信息产业领域把握先机和培育国民经济新增长点的关键。

此外，在相关软硬件产品研发和国际标准化形成方面，国内外的网络设备制造商、互联网巨头公司、云计算服务提供商和传统网络服务提供商等齐头并进，推出了日益丰富的 SDN 软硬件产品，并逐步部署和示范应用于数据中心内和数据中心间网络。随着相关国际标准化组织和产业联盟的推动，SDN 数据平面和控制平面的解耦分离，以及网络交换设备及控制器等的软硬件开放标准化，将推动国内外企业在不同技术层面的分工细化和市场竞争。国内企业及产品正面临着新一轮的机遇和挑战。

综上所述，如何继续加强数据中心网络的研究、增大相关项目资助力度、促进工业界与学术界的紧密联系，对于推动我国云计算和下一代互联网产业发展，并在国际新一轮 IT 技术革新浪潮中取得话语权，有至关重要的意义。

5 发展趋势与展望

云数据中心网络基础理论与关键技术研究有助于揭示云计算数据中心环境下网络设

计的科学规律与技术原理，为新一代云计算、大数据和互联网基础设施建设提供科学依据，推动网络技术、计算技术、数据科学等相关学科交叉融合发展，为我国战略性新兴产业发展提供具有自主知识产权的核心技术，对我国在国际新一轮信息技术革命浪潮中占据领先地位有非常重要的影响。

在本文介绍的云数据中心网络相关研究方向中，尽管当前学术界和工业界对于 SDN 的技术思路能否用于广域网还存在一些争议，但云数据中心网络被普遍认为是适用 SDN 的理想环境，这也被谷歌公司的实践以及中国电信与华为公司在 IDC 网络的合作所证明。由于 OpenFlow 协议尚存在不足，目前学术界正在积极探索其他可能的 SDN 架构和协议，这也为我国在 SDN 领域的技术创新和突破提供了机会。首先，尽管当前学术界和工业界已经对云数据中心网络互联拓扑展开了较为充分的研究，但在软件定义网络等新型网络架构下进行云数据中心网络拓扑的横向可扩展互联方面，还存在很大的研究空间，也具有更加重要的意义。第二，在目前提出的云数据中心网络路由和传输协议中，还缺乏对更为普遍使用的单播路由以及完全不同于 TCP 的新型传输协议的研究。为此，软件定义的网络框架将为云数据中心网络新型路由和传输协议设计提供更宽阔的技术选择空间。进一步，在软件定义的网络框架下如何充分利用可定制的网络功能，以更高效的方式配置虚拟网络并提供更安全的网络隔离，以及实现同时保障用户的网络带宽需求并优化网络资源利用率，将成为未来提高云计算服务用户体验的重要技术途径。与此同时，在软件定义的网络架构基础上，如何从多维度进行数据中心网络能耗的协同控制、交叉应用智能电网技术和清洁可再生能源优势，将成为未来大幅度降低云数据中心系统能耗、实现节能减排的绿色计算目标的重要手段，也是未来数年内的新兴研究热点。

参考文献

- [1] N Feamster, H Balakrishnan, J Rexford, etc. . The Case for Separating Routing From Routers[C]. In Proceedings of the ACM SIGCOMM workshop on Future Directions in Network Architecture. ACM, 2004: 5-12.
- [2] M Caesar, D Caldwell, N Feamster, etc. . Design and Implementation of a Routing Control Platform[C]. In NSDI, USENIX Association, 2005: 15-28.
- [3] M Casado, M J Freedman, J Pettit, etc. . Ethane: Taking Control of the Enterprise[J]. In ACM SIGCOMM Computer Communication Review. ACM, 2007, 37(4) : 1-12.
- [4] H Yan, D A Maltz, T S E Ng, etc. . Tesseract: A 4D Network Control Plane In NSDI, 2007.
- [5] N McKeown, T Anderson, H Balakrishnan, etc. . OpenFlow: Enabling Innovation in Campus Networks [C]. In ACM SIGCOMM Computer Communication Review, 2008, 38(2) : 69-74.
- [6] T Koponen, M Casado, N Gude, etc. . Onix: A Distributed Control Platform for Large- scale Production Networks[C]. In OSDI, Oct, 2010.
- [7] Z Cai, A L Cox, T S E Ng, Maestro: A System for Scalable OpenFlow Control[C]. In Tech. Rep. TR10-

- 08 , Rice University , 2010.
- [8] A R Curtis, J C Mogul, J Tourrilhes, etc . DevoFlow: Scaling Flow Management for High- Performance Networks[C]. In SIGCOMM-Computer Communication Review , 2011 , 41(4) : 254.
- [9] M Yu, J Rexford, M J Freedman, etc . Scalable Flow-Based Networking with DIFANE [J]. In ACM SIGCOMM Computer Communication Review. ACM , 2010 , 41(4) : 351-362.
- [10] M Al-Fares, A Loukissas, A Vahdat, A Scalable, Commodity Data Center Network Architecture[J]. In ACM SIGCOMM Computer Communication Review. ACM , 2008 , 38(4) : 63-74.
- [11] A Greenberg, J Hamilton, N Jain, etc . VL2: A Scalable and Flexible Data Center Network[J]. In ACM SIGCOMM Computer Communication Review. ACM , 2009 , 39(4) : 51-62.
- [12] G Wang, D Andersen, M Kaminsky, etc . e-Through: Part-time Optics in Data Centers[J]. In ACM SIGCOMM , 327-338 , Aug 2010.
- [13] K Chen, A Singla, A Singh, etc . OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility[C]. In NSDI USENIX Association , Apr 2012.
- [14] C Guo, H Wu, K Tan, etc . DCCell: A scalable and Fault-tolerant Network Structure for Data Centers[J]. In ACM SIGCOMM , 75-86 , Aug 2008.
- [15] C Guo, G Lu, D Li, etc . BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers[J]. In ACM SIGCOMM , 63-74 , Aug 2009.
- [16] H Wu, G Lu, D Li, etc . MDCube: A High Performance Network Structure for Modular Data Center Interconnection[J]. In ACM CoNext , 25-36 , Dec 2009.
- [17] D Li, M Xu, H Zhao, etc . Building Mega Data Center from Heterogeneous Containers[J]. In IEEE ICNP , 256-265 , Oct 2011.
- [18] D Li, Y Li, J Wu, S Su and J Yu, ESM: Efficient and Scalable Data Center Multicast Routing[C]. In IEEE/ACM Transactions on Networking , 20(3) : 944-955 , 2012.
- [19] D Li, H Cui, Y Hu, etc . Scalable Data Center Multicast Using Multi-class Bloom Filter[C]. In IEEE ICNP'11 , Vancouver , BC Canada. 2011 : 266-275.
- [20] V. Vasudevan, A Phanishayee, H Shah, etc . Safe and Effective Fine- grained TCP Retransmissions for Datacenter Communication[J]. In ACM SIGCOMM Computer Communication Review. ACM , 2009 , 39 (4) : 303-314.
- [21] M Alizadeh, A Greenberg, D Maltz, etc . Data Center TCP (DCTCP)[J]. In ACM SIGCOMM Computer Communication Review , 2010 , 40(4) : 63-74.
- [22] H Wu, Z. Feng, C Guo, etc . ICTCP: Incast Congestion Control for TCP in Data Center Networks[C]. In Proceedings of CoNext. ACM , 2010: 13. , Philadelphia.
- [23] C Raiciu, S Barre, C Pluntke, etc . Improving datacenter performance and robustness with multipath TCP [J]. In ACM SIGCOMM Computer Communication Review. ACM , 2011 , 41(4) : 266-277.
- [24] J Mudigonda, B Stiekes, P Yalagandula, etc . NetLord: A Scalable Multi-tenant Network Architecture for Virtualized Datacenters[C]. In ACM SIGCOMM'11 , 62-73 , Aug 2011.
- [25] D Eastlake, A Banerjee, D Dutt, etc . Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS [S]. RFC 6326 , Aug 2012.
- [26] M Mahalingam, D Dutt, K Duda, etc . VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks[S]. IETF draft , Feb 2013.

- [27] M Sridharan, K Duda, I Ganga, etc. . NVGRE: Network Virtualization using Generic Routing Encapsulation [S]. IETF draft, Feb 2013.
- [28] A Shieh, S Kandula, A Greenberg, etc. . Sharing the Data Center Network[C]. In USENIX NSDI'11, Apr 2011.
- [29] T Lam, S Radhakrishnan, A Vahdat, etc. . NetShare: Virtualizing Data Center Networks Across Services [J]. Technical Report, UCSD, 2010.
- [30] L Popa, A Krishnamurthy, S Ratnasamy, etc. . FairCloud: Sharing the Network in Cloud Computing[J]. In ACM SIGCOMM'12, 187-198, Aug 2012.
- [31] C Guo, G Lu, H Wang, etc. . SecondNet: A Data Center Network Virtualization Architecture with Bandwidth Guarantees[C]. In ACM CoNext'10, Nov 2010.
- [32] H Ballani, P Costa, T Karagiannis, etc. . Towards Predictable Datacenter Networks [J]. In ACM SIGCOMM'11, 242-253, Aug 2011.
- [33] D Xie, N Ding, Y C Hu, etc. . The Only Constant is Change: Incorporating Time-varying Network Reservations in Data Centers[J]. In ACM SIGCOMM'12, 199-210, Aug 2012.
- [34] V Jalaparti, H Ballani, P Costa, etc. . Bridging the Tenant-provider Gap in Cloud Services[J]. In ACM SOCC'12, Oct 2012.
- [35] C Liang, J Liu, L Luo, etc. . RACNet: A High-fidelity Data Center Sensing Network[J]. In ACM SenSys'09, Nov 2009.
- [36] B Weiss, H Truong, W Schott, etc. . Wireless Sensor Network for Continuously Monitoring Temperatures in Data Center[C]. IBM Research Report, 2011.
- [37] Lawrence Berkeley National Laboratory (LBLN)[J]. <http://www.lbl.gov/>.
- [38] R Mahdavi and W Tschudi, Wireless Sensor Network for Improving the Energy Efficiency of Data Centers [C]. LBLN Report for U. S General Services Administration, 2012.
- [39] M Gupta, S Grover, S Singh, A Feasibility Study for Power Management in LAN Switches[C]. In IEEE ICNP'04, 361-371, Oct 2004.
- [40] G Ananthanarayanan and R Katz. Greening the Switch[C]. HotPower'08, 7-11, Dec 2008.
- [41] M Gupta and S Singh. Using Low-Power Modes for Energy Conservation in Ethernet LANs [J]. In IEEE INFOCOM'07, 2451-2455, May 2007.
- [42] C Gunaratne, K Christensen, B Nordman. Managing Energy Consumption Costs in Desktop PCs and LAN Switches with Proxying, Split TCP Connections, and Scaling of Link Speed[J]. International Journal of Network Management, 2005, 15(5) : 297-310.
- [43] C Gunaratne, K Christensen, B Nordman, etc. Reducing the Energy Consumption of Ethernet with Adaptive Link Rate (ALR)[J]. IEEE Transactions on Computers, 57(4) : 448-461, Apr 2008.
- [44] B Heller, S Seetharaman and P Mahadevan. ElasticTree: Saving Energy in Data Center Networks[C]. In USENIX NSDI'10, Apr 2010.
- [45] G Cook. How Clean Is Your Cloud? [C]. In Greenpeace International Tech. Rep. , Apr 2012.
- [46] M Brown and J Renau. Rerack: Power Simulation for Data Centers with Renewable Energy Generation[J]. ACM GreenMetrics'11, 2011, 39(3) : 77-81.
- [47] C Ren, D Wang, B Urgaonkar, etc. . Carbon-Aware Energy Capacity Planning for Datacenters[J]. IEEE MASCOTS'12, 391-400, Aug 2012.

- [48] D Gmach, J Rolia, C Bash, etc. Capacity Planning and Power Management to Exploit Sustainable Energy [J]. IEEE/IFIP CNSM'10, 96-103, Oct 2010.
- [49] R Diaz, J Behr, M Tulpule. Energy Portfolio Simulation Considering Environmental and Public Health Impacts[J]. ACM EAIA'11, 38-45, Apr 2011.
- [50] Ymir Vigfusson, Hussam Abu-Libdeh, Mahesh Balakrishnan, Ken Birman, Yoav Tock. Dr. Multicast: Rx for Data Center Communication Scalability[C]. HOTNETS'08, Oct 2008.
- [51] M Al-Fares, S Radhakrishnan, B Raghavan, N Huang, A Vahdat. Hedera: Dynamic Flow Scheduling for Data Center Networks[C]. NSDI'10, pp. 19-19, 2010.
- [52] A Curtis, W Kim, P Yalagandula. Mahout: Low-overhead Datacenter Traffic Management Using End-host-based Elephant Detection[J]. IEEE INFOCOM, pp. 1629-1637, 2011.
- [53] A Dixit, P Prakash, Y Hu, R Kompella. On the Impact of Packet Spraying in Data Center Networks[J]. IEEE INFOCOM, pp. 2130-2138, 2013.
- [54] J Cao, R Xia, P Yang, C Guo, G Lu, L Yuan, Y Zheng, H Wu, Y Xiong, D Maltz. Per-packet Load-balanced, Low-latency Routing for Clos-based Data Center Networks[J]. ACM CoNext, pp. 49-60, 2013.
- [55] C Jiang, D Li, M Xu. Ltpp: An lt-code Based Transport Protocol for Many-to-one Communication in Data Centers[J]. Selected Areas in Communications, IEEE Journal on, vol. 32, no. 1, pp. 52-64, 2014.
- [56] Peng Cheng, Fengyuan Ren, Ran Shu, Chuang Lin. Catch The Whole Lot In An Action: Rapid Precise Packet Loss Notification In Data Centers[J]. ACM NSDI, pp. 17-28, 2014.
- [57] Xiaozhou Li, Michael J Freedman. Scaling IP Multicast On Datacenter Topologies[J]. ACM CoNEXT, pp. 61-72, 2013.
- [58] Siddhartha Sen, David Shue, Sunghwan Ihm, Michael J Freedman. Scalable, Optimal Flow Routing In Datacenters Via Local Link Balancing[J]. ACM CoNEXT, pp. 151-162, 2013.
- [59] Jiaxin Cao, Chuanxiong Guo, Guohan Lu, Yongqiang Xiong, Yixin Zheng, Yongguang Zhang, Yibo Zhu, Chen Chen. Datacast: A Scalable And Efficient Reliable Group Data Delivery Service For Data Centers[J]. ACM CoNEXT, pp. 37-48, 2012.
- [60] Cisco Data Center Infrastructure 2.5 Design Guide[OL]. http://www.cisco.com/application/pdf/en/us/guest/netsol/ns107/c649/ccmi-gration_09186a008073377d.pdf, 2007.
- [61] M Al-Fares, A Loukissas, A Vahdat. A Scalable, Commodity Data Center Network Architecture[J]. ACM SIGCOMM'08, 63-74, Aug 2008.
- [62] A Greenberg, J Hamilton, N Jain, etc. VL2: A Scalable and Flexible Data Center Network[J]. ACM SIGCOMM'09, 51-62, Aug 2009.
- [63] C Guo, H Wu, K Tan, etc. DCell: A Scalable and Fault-tolerant Network Structure for Data Centers [J]. ACM SIGCOMM'08, 75-86, Aug 2008.
- [64] K Ranachandran. 60GHz Data-Center Networking: Wireless = > Worryless[C]. Technical Report, NEC Laboratories America, 2008.
- [65] D Kedar, S Arnon. Urban Optical Wireless Communication Networks: The Main Challenges And Possible Solutions[J]. IEEE Communications Magazine, 2004, 42(5): S2-S7.
- [66] M Al-Fares, A Loukissas, A Vahdat. A Scalable, Commodity Data Center Network Architecture[C]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63-74.

- [67] M Niranjan, A Pamboris, N Farrington, H Nelson, et al. Portland: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric[C]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 39-50.
- [68] A Greenberg, J R Hamilton, N Jain, S Kandula, C Kim, P Lahiri, et al. VL2: A Scalable And Flexible Data Center Network[C]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 51-62.
- [69] C Wang, C R Wang, X Wang, D Jiang. Data Center Network Architecture Design towards Cloud Computing[C]. Journal of Computer Research and Development, 2012, 49(2): 286-293.
- [70] V Liu, D Halperin, A Krishnamurthy, T Anderson. F10: A Fault-Tolerant Engineered Network[C]. In Proc. of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2013). Lombard: USENIX, 2013.
- [71] D Abts, M R Marty, P Wells, P Klausler, H Liu. Energy Proportional Datacenter Networks[C]. ACM SIGARCH Computer Architecture News, 2010, 38(3): 338-347.
- [72] J Ahn, N Binkert, A Davis, M McLaren, R Schreiber. HyperX: Topology, Routing, Packaging of Efficient Large-Scale Networks [C]. In Proc. of the Conference on High Performance Computing Networking, Storage and Analysis. New York: ACM, 2009.
- [73] N Farrington, G Porter, S Radhakrishnan, H Bazzaz, V Subramanya, Y Fainman, A Vahdat. Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers[C]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 339-350.
- [74] G Wang, D Andersen, M Kaminsky, K Papagiannaki, T Ng, M Kozuch, M Ryan. C-Through: Part-Time Optics in Data Centers[C]. ACM SIGCOMM Computer Communication Review, 2010, 40(4): 327-38.
- [75] K Chen, A Singla, A Singh, K Ramachandran, L Xu, Y Zhang, X Wen, Y Chen. OSA: An Optical Switching Architecture For Data Center Networks With Unprecedented Flexibility[C]. In Proc. of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2012). San Jose: USENIX, 2012.
- [76] H Wang, Y Xia, K Bergman, T Ng, S Sahu, K Sripanidkulchai. Rethinking The Physical Layer Of Data Center Networks Of The Next Decade: Using Optics To Enable Efficient * - Cast connectivity[C]. ACM SIGCOMM Computer Communication Review, 2013, 43(3): 52-58.
- [77] D Guo, T Chen, D Li, Y Liu, G Chen. Expandable And Cost-Effective Network Structures For Data Centers Using Dual-port Servers[C]. IEEE Transactions on Computers, 2013, 62(7): 1303-1317.
- [78] D Li, C Guo, H Wu, K Tan, Y Zhang, S Lu. FiConn: Using Backup Port For Server Interconnection In Data Centers [C]. In Proc. of the 28th IEEE International Conference on Computer Communications (INFOCOM 2009), Rio de Janeiro: IEEE 2009. 2276-2285.
- [79] C Guo, G Lu, D Li, H Wu, X Zhang, Y Shi, C Tian, Y Zhang, S Lu. BCube: A High Performance, Server-centric Network Architecture For Modular Data Centers [C]. ACM SIGCOMM Computer Communication Review, 2009, 39(4): 63-74.
- [80] H Abu-Libdeh, P Costa, A Rowstron, G O'Shea, A Donnelly. Symbiotic Routing In Future Data Centers [C]. ACM SIGCOMM Computer Communication Review, 2011, 41(4): 51-62.
- [81] X Liu, S Yang, L Guo, S Wang, H Song. Snowflake: A New-Type Network Structure of Data Center [C]. Chinese Journal of Computers, 2011, 34(1): 76-85.
- [82] H Wu, G Lu, D Li, C Guo, Y Zhang. MDCube: A High Performance Network Structure For Modular

- Data Center Interconnection[C]. In Proc. of the 5th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT 2009). Rome: ACM, 2009. 25-36.
- [83] D Li, M Xu, H Zhao, X Fu. Building Mega Data Center From Heterogeneous Containers[C]. In Proc. of the 19th IEEE International Conference on Network Protocols (ICNP 2011), Vancouver: IEEE, 2011. 256-265.
- [84] H Wu, Z Feng, C Guo, etc. . ICTCP: Incast Congestion Control for TCP in Data Center Networks[C]. ACM CoNext 2010, Philadelphia.
- [85] C Guo, G. Lu, H Wang, etc. . SecondNet: A Data Center Network Virtualization Architecture With Bandwidth Guarantees[C]. ACM CoNext 2010, Philadelphia, PA
- [86] G Lu, C Guo, Y Li, etc. . ServerSwitch: A Programmable and High Performance Platform for Data Center Networks[C]. USENIX NSDI2011, Berkeley, CA
- [87] J Shin, B Wong, E Sizer. Small-world datacenters[C]. In Proc. of the 2nd ACM Symposium on Cloud Computing (SOCC 2011). Cascais: ACM, 2011.
- [88] A Singla, C Hong, L Popa, G Brighten. Jellyfish: Networking Data Centers Randomly[C]. In Proc. of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2012), San Jose: USENIX, 2012.
- [89] L Gyarmati and T Trinh. Scafida: A Scale-Free Network Inspired Data Center Architecture [C]. ACM SIGCOMM Computer Communication Review, 2010, 40(5): 4-12.
- [90] J Shin, E Sizer, H Weatherspoon, K Darko. On The Feasibility of Completely Wireless Datacenters[C]. In Proc. of the 8th ACM/IEEE Symposium on Architectures for Networking and Communications Systems (ANCS 2012). Austin: ACM/IEEE, 2012. 3-14.
- [91] X Zhou, Z Zhang, Y Zhu, Y Li, K Saipriya, V Amin, B Zhao, H Zheng. Mirror Mirror On The Ceiling: Flexible Wireless Links For Data Centers [C]. ACM SIGCOMM Computer Communication Review, 2012, 42(4): 443-454.
- [92] N Hamedazimi, H Gupta, V Sekar, R Samir. Patch Panels In The Sky: A Case For Free-Space Optics In Data Centers[C]. In Proc. of the 12th ACM SIGCOMM Workshop on Hot Topics in Networks (HotNets 2013). Hong Kong: ACM, 2013.
- [93] D Li, J Yu, J Yu, etc. . Exploring Efficient and Scalable Multicast Routing in Future Data Center Networks[C]. IEEE INFOCOM 2011, Shanghai, China.
- [94] D Li, H Cui, Y Hu, etc. . Scalable Data Center Multicast Using Multi-Class Bloom Filter[C]. IEEE ICNP 2011, Vancouver, BC Canada.
- [95] J Zhang, F Ren, C Lin. Modeling And Understanding TCP Incast In Data Center Networks[C]. IEEE INFOCOM 2011, Shanghai, China.
- [96] C Jiang, D Li and M Xu. A Coding-based Approach to Mitigating TCP Incast in Data Center Network[C]. ICDCS Workshop on DCPerf 2012, Macao, China.
- [97] Y Cui, H Wang, X Cheng, etc. . Wireless Data Center Networking[C]. IEEE Wireless Communications, 2011.
- [98] Y Cui, H Wang, X Cheng. Channel Allocation in Wireless Data Center Networks[C]. IEEE INFOCOM 2011, Shanghai, China.
- [99] F Ren, W Jiang. Phase PlaneAnalysis of Congestion Control in Data Center Ethernet Networks[C]. ICDCS

2010, Genoa, Italy.

- [100] V Vasudevan, A Phanishayee, H Shah, etc. . Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication[C]. ACM SIGCOMM 2009, Barcelona, Spain.
- [101] Y Shang, D Li, M Xu. Energy-aware Routing in Data Center Network[C]. ACM SIGCOMM Workshop on Green Networking 2010, New Delhi, India.
- [102] Y Shang, D Li, M Xu. A Comparison Study of Energy Proportionality of Data Center Network Architectures [C]. ICDCS Workshop on DCPerf2012, Macao, China.
- [103] D Li, Y Yu, K Li. SODA: Enhancing the Data Plane Functionality of Software Defined Networking[C]. Open Network Summit 2014, Santa Clara, CA
- [104] D Li, Y Shang, C Chen. Software Defined Green Data Center Network with Exclusive Routing[C]. IEEE INFOCOM 2014, Toronto, Canada.
- [105] Cui Yong, et al. . Data Centers As Software Defined Networks: Traffic Redundancy Elimination With Wireless Cards At Routers[C]. Selected Areas in Communications, IEEE Journal on 31.12 (2013): 2658-2672.
- [106] G Xie, P He, H Guan, etc. . PEARN: A Programmable Virtual Router Platform, IEEE Communication Magazine[C]. Special Issue on Future Internet Architectures: Design and Deployment Perspectives, July, 2011.
- [107] L Luo, G Xie, Y Xie, L Mathy, K Salamatian. A Hybrid Hardware Architecture for High-Speed IP Lookups and Fast Route Updates [J]. IEEE/ACM Transactions on Networking 01/2014; 22 (3): 957-969.
- [108] G Xie, Y Sun, Y Zhang, etc. . Service-Oriented Future Internet Architecture (SOFIA) [C]. IEEE Infocom/Poster, Shanghai, China, April 2011.
- [109] Q Wu, Z Li, G Xie. CodingCache: Multipath-Aware CCN Cache With Network Coding[C]. In Proc. the 3rd ACM SIGCOMM workshop on Information-centric networking, 2013.
- [110] J Guo, F Liu, D Zeng, J Lui, H Jin. A Cooperative Game Based Allocation for Sharing Data Center Networks[C]. in Proc. of IEEE INFOCOM, Italy, April, 2013.
- [111] J Guo, F Liu, H Tang, Y Lian, H Jin, J Lui. Falloc: Fair Network Bandwidth Allocation in IaaS Datacenters Via a Bargaining Game Approach [C]. in Proc. of IEEE ICNP, Goettingen, Germany, October, 2013.
- [112] J Guo, F Liu, X Huang, J Lui, M Hu, Q Gao, H Jin. On Efficient Bandwidth Allocation for Traffic Variability in Datacenters[C]. in Proc. of IEEE INFOCOM, Toronto, April, 2014.
- [113] F Xu, F Liu, H Jin, A Vasilakos. Managing Performance Overhead of Virtual Machines in Cloud Computing: A Survey, State of Art and Future Directions [C]. Proceedings of the IEEE, vol. 102, no. 1, Jan. 2014.
- [114] F Xu, F Liu, L Liu, H Jin, B Li, B Li. iAware: Making Live Migration of Virtual Machines Interference-Aware in the Cloud[C]. IEEE Transactions on Computers, 2014.
- [115] F Liu, Z Zhou, H Jin, B Li, Baochun Li, Hongbo Jiang. On Arbitrating the Power-Performance Tradeoff in SaaS Clouds[C]. accepted by IEEE Transactions on Parallel and Distributed Systems, 2014.
- [116] 邓维, 刘方明, 金海, 李丹. 云计算数据中心的新能源应用: 研究现状与趋势[J]. 计算机学报, 2013, 36(3): 582-588.

- [117] W Deng, F Liu, H Jin, C Wu. SmartDPSS: Cost-Minimizing Multi-source Power Supply for Datacenters with Arbitrary Demand[C]. in Proc. of ICDCS, Philadelphia, USA, July, 2013.
- [118] Z Zhou, F Liu, Y Xu, R Zou, H Xu, J Lui, H Jin. Carbon-aware Load Balancing for Geo-distributed Cloud Services[C]. in Proc. of IEEE MASCOTS, August, San Francisco, USA, 2013.
- [119] Z Zhou, F Liu, B Li, B Li, H Jin, R Zou, Z Liu. Fuel Cell Generation in Geo- Distributed Cloud Services: A Quantitative Study[C]. in Proc. of ICDCS, Madrid, Spain, July, 2014.
- [120] C Hu, J Yang, H Zhao, J Lu. Design of All Programmable Innovation Platform for Software Defined Networking [C]. Open Networking Summit (ONS) 2014, Research Track, Santa Clara, Mar. 2-5, 2014.
- [121] C Hu, J Yang, S Deng, Z Gong, H Zhao. DesktopDC: Setting All Programmable Data Center Networking Testbed on Desk[C]. SIGCOMM 2014, poster, Chicago, Aug. 17-22, 2014.
- [122] T Huang, C Rong, Y Tang, C Hu, J Li, P Zhang. VirtualRack: Bandwidth- Aware Virtual Network Allocation for Multi-Tenant Datacenters[C]. in the Proceeding of IEEE ICC 2014, Sydney, Australia, Jun. , 2014.
- [123] X Wen, K Chen, Y Chen, Y Liu, Y Xiang, C Hu. VirtualKnotter: Online Virtual Machine Shuffling for Congestion Resolving in Virtualized Datacenter[C]. IEEE ICDCS 2012, Jun. 18-21, Macau, 2012.
- [124] C Hu, M Yang, K Zheng, K Chen, X Zhang, B Liu, X Guan. Automatically Configuring the Network Layer of Data Centers for Cloud Computing[C]. IBM Journal of Research and Development, Vol. 5, No. 6, pp. 3:1-3:10, 2011.
- [125] X Ma, C Hu, K Chen, C Zhang, H Zhang, K Zheng, Y Chen, X Sun. Error Tolerant Address Configuration for Data Center Networks with Malfunctioning Devices[J]. IEEE ICDCS 2012, Jun. 18-21, Macau, 2012.
- [126] K Chen, C Hu, X Zhang. Survey on Routing in Data Centers: Insights and Future Directions[J]. IEEE Network, 25(4) : 6-10. 2011.
- [127] D Guo, L Luo, X Zhou, J Wu, X Luo. DCube: A Family of High Performance Modular Data Centers Using Dual-Port Servers[J]. Accepted to appear at Elsevier Journal of computer communication, 2014.
- [128] Z Ding, D Guo, X Liu, etc. . A MapReduce- supported Network Structure for Data Centers [J]. to appear at Concurrency and Computation: Practice and Experience, 2011.
- [129] D Guo, M Li, H Jin, etc. . Aggregating Data Transfers in Data Centers[C]. National University of Defense Technology, Changsha, Hunan, China, Tech. Rep. , Jul. 2012
- [130] D Guo, J Xie, X Zhou, X Zhu, W Wei, X Luo. Exploiting Efficient and Scalable Shuffle Transfers in Future Data Center Networks[C]. accepted to appear in IEEE Transactions on Parallel and Distributed Systems, 2014.
- [131] D Zhang, D Guo, F Chen, etc. . TL- Plane- Based Multi- Core Energy- Efficient Real- Time Scheduling Algorithm for Sporadic Tasks[J]. ACM Transactions on Architecture and Code Optimization (TACO) , 2012, 8(4) : 47.
- [132] J Xie, D Guo. Controller Deployment in Data Centers[C]. National University of Defense Technology, Changsha, Hunan, China, Tech. Rep. , Jul. 2014.
- [133] R Tu, C Zhou, J Zhao, X Wang. SDN Middle Box for Data Centers[S]. IETF draft-tu-sdnrg-middle-box-00, July 4, 2014.

- [134] X Zhang, H Wang, Q Gong, X Wang. Decluster: A Complex Network Model-Based Data Center Network Topology [J]. *The Journal of Supercomputing* (Springer), May 2014, DOI: 10.1007/s11227-014-1232-8.
- [135] 吴舢, 屠仁龙, 王新. SDN 在 Web 服务负载均衡与性能优化中的应用[C]. 第三届中国互联网学术会议, 上海, 2014 年 7 月.
- [136] R Tu, X Wang, Y Yang. Energy-Saving Model for SDN Data Centers[J]. *The Journal of Supercomputing* (Springer), May 2014, DOI: 10.1007/s11227-014-1237-3.
- [137] Y Li, H Wang, P Zhang, etc. D4D: Inter- Datacenter Bulk Transfers with ISP Friendliness [C]. to appear in IEEE Clsuter 2012, Beijing, China.
- [138] P Zhang, H Wang, S Cheng. Shrinking MTU to Improve Fairness Among TCP Flows in Data Center Networks[C]. ICCSN 2011, May 2011.
- [139] Y Ma, H Wang, J Dong, etc. ME2 Efficient Live Migration of Virtual Machine With Memory Exploration and Encoding[C]. IEEE Clsuter 2012, Beijing, China.
- [140] P Zhang, H Wang, J Li, J Dong, Y Li, S Cheng. SmartShuffle: Managing Online Virtual Machine Shuffle in Virtualized Data Centers[C]. In Proc. the 34rd IEEE International Conference on Distributed Computing Systems (ICDCS 2013), July 08-11, 2013, Philadelphia, USA.
- [141] J Dong, H Wang, X Jin, Y Li, P Cheng. Virtual Machine Placement for Improving Energy Efficiency and Network Performance in IaaS Cloud [C]. In Proc. the 34rd IEEE International Conference on Distributed Computing Systems (ICDCS 2013), July 08-11, 2013, Philadelphia, USA.
- [142] J Dong, X Jin, H Wang, Y Li, P Zhang, S Cheng. Energy-Saving Virtual Machine Placement in Cloud Data Center[C]. In Proc. the 13th IEEE/ACM CCGrid, July, 2013, Delft, Netherlands.
- [143] Y Li, H Wang, J Dong, J Li, S Cheng. Operating Cost Reduction for Distributed Internet Data Centers [C]. In Proc. the 13th IEEE/ACM CCGrid, July, 2013, Delft, Netherlands.
- [144] K Li, Y Mu, K Li, G. Min. Exchanged Crossed Cube: A Novel Interconnection Network for Parallel Computation[J]. *IEEE Transactions on Parallel and Distributed Systems*, 24(11), 2211-2219, 2013.
- [145] Y Li, H Qi, Z Li, K Li. The Exchanged Folded Hypercube [C]. The 15th IEEE International Conference on High Performance Computing and Communications (HPCC), 2013.
- [146] Y Shen, S Panwar, H Chao. SQUID: A Practical 100% Throughput Scheduler for Crosspoint Buffered Switches[J]. *IEEE/ACM Transactions on Networking*, 18(4), 1119-1131, 2010.
- [147] 王聪, 王翠荣, 王兴伟, 蒋定德. 面向云计算的数据中心网络体系结构设计[J]. 计算机研究与发展, 2012, 49(2): 286-293.
- [148] <http://www.meshsr.com/product>.
- [149] Ballani, Hitesh, Keon Jang, Thomas Karagiannis, Changhoon Kim, Dinan Gunawardena, Greg O'Shea. Chatty Tenants and the Cloud Network Sharing Problem[C]. In NSDI, 171-184. 2013.
- [150] Jeyakumar, Vimalkumar, Mohammad Alizadeh, David Mazieres, Balaji Prabhakar, Changhoon Kim, Albert Greenberg. EyeQ: Practical Network Performance Isolation at the Edge[J]. REM 1005, no. A1 (2013) : A2.
- [151] Popa, Lucian, Praveen Yalagandula, Sujata Banerjee, Jeffrey C Mogul, Yoshio Turner, Jose Renato Santos. ElasticSwitch: practical work-conserving bandwidth guarantees for cloud computing [J]. In Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM, 351-362. ACM, 2013.

- [152] The Road to SDN: An Intellectual History of Programmable Networks
- [153] McKeown, T Anderson, H Balakrishnan, G Parulkar, L Peterson, J Rexford, S Shenker, J Turner. OpenFlow: Enabling Innovation in Campus Networks [J]. ACM SIGCOMM Computer Communications Review, Apr. 2008.
- [154] M Casado, M J Freedman, J Pettit, J Luo, N McKeown, S Shenker. Ethane: Taking control of the enterprise[C]. In ACM SIGCOMM '07, 2007.
- [155] K Greene. TR10: Software-defined networking[J/OL]. MIT Technology Review, March/April 2009. <http://www2.technologyreview.com/article/412194/tr10-software-defined-networking/>
- [156] ONF Member Listing[OL]. <https://www.opennetworking.org/membership/member-listing>
- [157] Software-Defined Networking: The New Norm for Networks[C]. ONF White Paper April 13, 2012.
- [158] 张卫峰. 深入思考 SDN 的核心本质: 从 SDN = OpenFlow 回到软件定义网络[OL]. <http://www.csdn.net/article/2014-02-13/2818409-SDN>.
- [159] GENI: Global Environment for Network Innovations[OL]. <http://www.geni.net/>.
- [160] B Anwer, M Motiwala, M bin Tariq, N Feamster. SwitchBlade: A Platform for Rapid Deployment of Network.
- [161] M Dobrescu, N Egi, K Argyraki, B- G. Chun, K Fall, G Iannaccone, A Knies, M Manesh, S Ratnasamy. RouteBricks: Exploiting parallelism to scale software routers [C]. In Proc. 22nd ACM Symposium on Operating Systems Principles (SOSP), Big Sky, MT, Oct. 2009.
- [162] P Bosshart, G Gibb, H Kim, G Varghese, N McKeown, M Izzard, F Mujica, M Horowitz. Forwarding Metamorphosis: Fast Programmable Match- Action, Processing in Hardware for SDN [C]. In ACM SIGCOMM, Aug. 2013.
- [163] A Greenberg, G Hjalmtysson, D A Maltz, A Myers, J Rexford, G Xie, H Yan, J Zhan, H Zhang. A Clean Slate 4D Approach To Network Control And Management [J]. ACM SIGCOMM Computer Communications Review, 35(5): 41-54, 2005.
- [164] M Caesar, N Feamster, J Rexford, A Shaikh, J van der Merwe. Design and Implementation of A Routing Control Platform[C]. In Proc. 2nd USENIX NSDI, Boston, MA, May 2005.
- [165] T V Lakshman, T Nandagopal, R Ramjee, K Sabnani, T Woo. The SoftRouter Architecture[C]. In Proc. 3rd ACM Workshop on Hot Topics in Networks (Hotnets-III), San Diego, CA, Nov. 2004.
- [166] J van der Merwe, A Cepleanu, K D'Souza, B Freeman, A Greenberg, et al. Dynamic Connectivity Management With an Intelligent Route Service Control Point [C]. In ACM SIGCOMM Workshop on Internet Network Management, Oct. 2006.
- [167] P Verkaik, D Pei, T Scholl, A Shaikh, A Snoeren, J van der Merwe. Wrestling Control From BGP: Scalable Fine-grained Foute Control[C]. In Proc. USENIX Annual Technical Conference, June 2007.
- [168] Jeffrey C Mogul, Paul Congdon. Hey, You Darned Counters! Get off My ASIC! [C]. HotSDN'12 Proceedings of the first workshop on Hot topics in software defined networks, Pages 25-30.
- [169] Guohan Lu, Rui Miao, Yongqiang Xiong, Chuanxiong Guo. Using CPU as a Traffic Co-processing Unit in Commodity Switches[C]. HotSDN '12: Proceedings of the first workshop on Hot topics in software defined networks, Pages 31-36.
- [170] Minlan Yu, Jennifer Rexford, Michael J Freedman, Jia Wang. Scalable Flow-based Networking with DIFANE[C]. SIGCOMM'10: Proceedings of the ACM SIGCOMM 2010 conference.

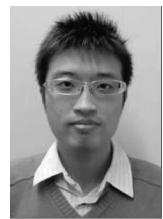
- [171] T Koponen, M Casado, N Gude, J Stribling, L Poutievski, M Zhu, R Ramanathan, Y Iwata, H Inoue, T Hama, S Shenker. Onix: A Distributed Control Platform for Large-Scale Production Networks [C]. In OSDI, volume 10, pages 1-6, 2010.
- [172] ON. Lab. ONOS: Open Network Operating System[OL]. <http://tinyurl.com/pjs9eyw>.
- [173] N Gude, T Koponen, J Pettit, B Pfaff, M Casado, N McKeown, S Shenker. NOX: Towards An Operating System For Networks [J]. ACM SIGCOMM Computer Communication Review, 38 (3): 105-110, July 2008.
- [174] Amin Tootoonchian, Yashar Ganjali. HyperFlow: A Distributed Control Plane for OpenFlow[C]. INM/WREN'10: Proceedings of the 2010 Internet Network Management Conference on Research on Enterprise Networking.
- [175] Zafar Ayyub Qazi, Cheng-Chun Tu, Luis Chiang, Rui Miao, Vyas Sekar, Minlan Yu. SIMPLE-fying Middlebox Policy Enforcement Using SDN[C]. SIGCOMM '13: Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM.
- [176] Seyed Kaveh Fayazbakhsh, Vyas Sekar, Minlan Yu, Jeffrey C Mogul. FlowTags: Enforcing Network-Wide Policies in the Presence of Dynamic Middlebox Actions[C]. NSDI'14: Proceedings of the 11th USENIX Conference on Networked Systems Design and Implementation.
- [177] S Jain, A Kumar, S Mandal, J Ong, L Poutievski, A Singh, S Venkata, J Wanderer, J Zhou, M Zhu, J Zolla, U Hlzle, S Stuart, A Vahdat. B4: Experience with a globally deployed software defined WAN[C]. SIGCOMM'13: Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM.

作者简介

李丹男，现任清华大学计算机系副教授，主要从事互联网、数据中心网络、云计算方向的研究工作。2008年1月获清华大学计算机科学与技术博士学位。2008年1月至2010年2月，任微软亚洲研究院副研究员。2010年3月加入清华大学计算机系工作至今。担任国家973计划（青年科学家专题）“软件定义的云数据中心网络基础理论与关键技术”项目负责人，主持或参与国家863项目、自然科学基金项目、国家242项目等10多个项目。担任国际学术期刊 IEEE Transactions on Computers 编委，国际学术会议 IEEE LANMAN 2014 程序委员会主席、IEEE LANMAN 2015 大会主席、IEEE ICNP poster/demo 程序委员会主席、IEEE INFOCOM 2010-2015 程序委员会委员。在 IEEE/ACM Transactions on Networking、IEEE Journal on Selected Areas in Communications、IEEE Transactions on Computers、ACM SIGCOMM、IEEE ICNP、IEEE INFOCOM 等网络领域的著名期刊和会议上发表论文50余篇。申请美国专利6项、中国专利20余项。



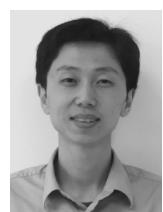
刘方明 博士，CCF 互联网专业委员会常委、CCF YOCSEF 委员和武汉分论坛学术委员。先后于清华大学获学士学位、于香港科技大学获博士学位、于加拿大多伦多大学任访问学者，现任华中科技大学副教授、博导，研究方向包括云计算与绿色计算、数据中心与软件定义网络 SDN、互联网“云—端”融合与对等计算模式的大规模分布式系统。担任国家 973 计划（青年科学家专题）“软件定义的云数据中心网络基础理论与关键技术”子课题负责人，并主持或参与国家 863 项目、自然科学基金项目、重点项目等多个项目，在包括国际著名期刊《Proceedings of the IEEE》、《IEEE Journal on Selected Areas in Communications》、《IEEE Transactions on Parallel & Distributed Systems》、《IEEE Transactions on Computers》以及顶尖学术会议 INFOCOM、ICDCS、ICNP、ACM NOSSDAV、ACM e-Energy 等上发表论文 40 余篇，获 IEEE GLOBECOM 等多项国际学术会议最佳论文奖，所设计实现的云存储与网络硬盘系统资源调度技术获大规模实际应用部署，用户超过 300 万。入选湖北省“楚天学者”、微软亚洲研究院 MSRA “铸星计划”、获华中科技大学“学术新人奖”和“华中学者”称号。担任著名 SCI 刊物《IEEE Network Magazine》和《IEEE Systems Journal》特邀编委、SCIE 期刊《Frontiers of Computer Science》首届青年 Associate Editor，担任 CCF A 类国际学术会议 IEEE INFOCOM 2013-2015、ACM Multimedia 2014、以及 ICNP 2014 等 10 余个大会程序委员会委员、IEEE LANMAN 2014 大会宣传主席和第九届绿色普适计算与云计算国际学术会议 GPC 2014 大会程序委员会主席。



郭得科 博士，CCF 开放系统专委委员、CCF 互联网专委委员、CCF 网络与数据通信专委委员。国防科学技术大学信息系统与管理学院副教授，国家自然科学基金优秀青年基金获得者，教育部新世纪优秀人才计划入选者，香港科技大学和南洋理工大学访问学者。研究方向包括分布式系统资源管理、网络系统架构、软件定义的数据中心网络等。在 IEEE TC、IEEE TPDS、IEEE TKDE、ACM TACO、ACM TOIT、INFOCOM、COMNET、COMCOM 等著名学术期刊和会议上发表 70 多篇论文。主持国家 973 青年科学家专项分课题、国家自然科学基金优秀青年基金等 10 多项国家级课题。在 Dynamic Bloom Filters 领域的研究成果得到工业界高度评价，被著名的云计算和大数据处理开源平台 Hadoop 集成实现。2011 年获湖南省优秀博士学位论文。



何 源 2003 年于中国科学技术大学获工学学士学位，2006 年于中国科学院软件研究所获工学硕士学位，2010 年于香港科技大学获得博士学位。现任清华大学软件学院特聘副研究员，国家自然科学基金优秀青年基金获得者。研究方向涉及分布式系统、无线和传感器网络、云计算、移动计算等，作为迄今国际上规模最大的室外无线传感器网络系统 GreenOrbs 绿野千传的项目工作组组长，全程参与了该系统的研发、部署和项目管理工作。担任国



际学术期刊《Ad Hoc and Sensor Wireless Networks》等的编委、学术会议 IEEE DCOSS、MASS、ICPADS、ICC、GLOBECOM 等的程序委员会委员，在《IEEE/ACM Transactions on Networking》、《IEEE Transactions on Parallel and Distributed Systems》、《IEEE Transactions on Mobile Computing》等国际学术期刊和 IEEE INFOCOM、ACM SenSys、IEEE ICNP、IEEE RTSS 等国际学术会议上已发表论文 70 余篇，获得 IEEE ICPADS 2010 年会唯一最佳论文奖。E-mail：he@greenorbs.com。

陈贵海 1984 年获南京大学士学位，1987 年获东南大学硕士学位，1997 年获香港大学博士学位。曾任教于日本九州工业大学、澳大利亚昆士兰大学和美国韦恩州立大学。主要研究方向包括无线网络、对等计算、高性能计算机系统结构、海量数据处理、组合数学等。已发表论文 300 余篇，其中国际刊物及国际会议论文 200 余篇，包括 IEEE TON、IEEE TC、IEEE TPDS 等国际知名刊物以及 MOBICOM、MOBIHOC、INFOCOM、ICNP、ICDCS、HPCA 等国际一流会议文章。被引用 7000 余次，单篇论文最高引用 600 余次。担任国际学术会议程序委员或大会主席 60 余次，现任中国计算机学会开放系统专委会主任。曾获多种奖励，包括教育部高校青年教师奖、国家自然科学基金委员会项目特优评价、中创软件人才奖、国家杰出青年科学基金、国务院政府特殊津贴等。



未来互联网体系结构研究现状与发展趋势

CCF 互联网专业委员会

罗洪斌¹ 胡宇翔² 毕军³ 李振宇⁴

¹北京交通大学电子与信息工程学院，北京

²中国人民解放军信息工程大学，郑州

³清华大学网络科学与网络空间研究院网络体系结构和 IPv6 研究室，北京

⁴中国科学院计算技术研究所，北京

摘要

随着用户和应用规模的不断扩大，现有互联网逐渐暴露出可扩展性差、移动性支持差、安全性差、能耗大等诸多问题，急需创建全新的未来互联网体系与机制，以适应经济社会发展的迫切需要。本文分析了未来互联网体系结构的国际发展现状和趋势，对我国多家单位在此方面的进展进行了介绍，并进行了简单对比与展望。

关键词：未来互联网，互联网体系结构，信息中心网络，内容中心网络

Abstract

With the rapid increase in the number of users and applications, the current Internet architecture exposes many deficits such as poor routing scalability, inefficient support for mobility, poor security, high energy consumption. Therefore, it is urgent to design novel Internet architecture and corresponding mechanisms, so as to better serve our society. The report makes an overview on the current situation and trends of future Internet architectures proposed around the world, and introduced some representative future Internet architectures proposed by China. We also make a simple comparison on these efforts and outline some critical remaining research challenges.

Keywords: Future Internet, Internet architecture, information- centric networking, content-centric networking.

1 引言

随着科学技术的发展，信息已经成为当今社会发展的巨大推动力。互联网作为信息的一个成功载体，已经渗透到政治、经济、文化、教育、卫生等人类社会生活的方方面面，成为人们日常生活不可缺少的一部分。然而，现有互联网是在上世纪 70 年代设计的，采用相对“静态”和“僵化”的设计思想，在其发展过程中表现出各种各样的原始

设计缺陷与不足，比如可扩展性差、移动性支持差、安全性差、资源利用率低、能耗大等诸多问题，难以满足未来网络“高速”、“高效”、“海量”、“泛在”等通信需求。因此，世界各国近年来相继开展了未来信息网络体系理论的相关研究，并针对其发展制定了相应的宏观决策，力图抢占未来信息网络领域的制高点。

例如，OpenSig（1996年）、Active Networks（1996年）、IEEE1520（1998年）、ForCES（2002年）均是这方面研究的最早开拓者；国际上也出现了大量针对后IP时代的新型网络基本体系结构及关键技术的研究，比较典型的如美国NSF资助的GENI（Global Environment for Network Innovation）计划^[1]、FIND（Future Internet Network Design）计划^[2]、欧盟FP7中下一代网络计划^[3]、ITU-T的NGN计划^[4]、日本的AKARI计划^[5]、韩国的下一代网络BCN（Broadband Convergence Network）计划^[6]、中国科技部“863”计划“新一代高可信网络”^[7]等。这些研究计划均试图以革新或演变方式改变已有网络系统设计，力图满足人类对信息网络的各种需求。

本报告将对国内外近年来在未来互联网体系结构方面的研究进展进行综述，比较国内外的研究进展，并预测未来的发展趋势。

2 国际研究现状

由于现有互联网存在可扩展性差、移动性支持差、安全性差、资源利用率低、能耗大等诸多不足，国际上对未来互联网体系结构的研究也各有侧重。概括起来，可以分为开放可编程网络、面向服务网络、内容中心网络、面向移动性网络以及典型的网络试验床等。

1) 开放可编程网络体系结构包括：较早提出的ForCES体系，以及近几年提出的软件定义网络（Software-Defined Network，SDN）。

2) 以服务为中心构建未来互联网络能够改变传统网络面向不同业务需求时只完成“傻瓜式”传输的窘境，实现传统互联网向商务基础设施、社会文化交流基础架构等新角色的转型。服务（Service）涵盖了传输和应用等，它利用数据资源、计算资源、存储资源、传输资源，完成对信息的高效、安全计算、存储及传输。面向服务的新型网络体系结构（Service Oriented Architecture，SOA）借鉴了软件设计中面向服务的架构设计、面向对象的模块化编程思想，将服务作为基本单元设计未来网络的各种功能，包含了对服务进行命名、注册、发布、订阅、查找、传输等各种功能的设计，以此满足未来新型网络的管理、传输、计算等需求。

3) 面向内容的网络体系结构是其中重要的研究方向之一，典型代表有NDN、DONA、PSIRP和NetInf。

4) 传统TCP/IP网络体系结构存在IP地址语义过载的问题，即IP地址既扮演着节点标识符的角色，又扮演着节点定位符的角色。IP地址语义过载问题会影响计算机网络对移动性的支持，限制核心路由的扩展性，降低现有安全机制的效能，还会限制若干新

技术的发展。

针对 IP 地址存在的缺陷，人们意识到网络标识不能够简单地定义为位置标志，而应当能够严格区分固定标识和可变标识，从而建立网络实体标识和网络位置标识两套标识系统。

国际 IAB 组织提出通过引入两个名字空间来分别表示节点的标识和位置，即所谓的“Locator/Identifier Split”，解决 IP 地址语义过载问题。

为了实现“Locator/Identifier Split”，需要引入 Locator 和 Identifier 两套名字空间，并完成两套名字空间之间的转换。这一机制的引入也对网络体系结构提出了新的要求，在映射服务、地址前缀聚合、地址绑定机制及移动性支持等方面提出了重大挑战。

身份与位置分离之后，需要在网络中引入一个映射系统，负责边缘网络所使用的地址与核心网络地址之间的映射。典型的机制有 APT、LISP、IvIP、TRRP、Six/One、Six/One Router。

5) 目前业界已经提出了很多创新型网络体系。但缺乏对这些新型网络体系进行真实性试验的环境，因此迫切需要具有可控和真实的网络试验平台。当前，欧盟的 FIRE 计划、美国的 FIND/GENI 计划、日本的 AKARI 计划均构建了试验床来测试和验证它们的解决方案。

2.1 开放可编程网络

开放架构网络的研究开始于 1996 年，是基于 3 种不同的开放架构的实现思想进行的：

- 1) 基于开放信令 (OpenSig)^[3]的思想。
- 2) 基于动态代码的主动网络 (Active Network)^[4]思想。
- 3) 通过资源预留的 Virtual Network^[5]思想。

上述 3 方面思想对开放架构网络研究具有一定的互补性，其共同目标都是实现网络的开放可编程性。然而，几乎所有开放可编程网络都基本采用了控制面 (Control Plane) 和数据面 (Data Plane) 分离的基本体系结构。

2.1.1 ForCES 体系结构

在上述众多与开放可编程网络有关的研究中，由于得到 IETF、ITU、NPF 等多家标准制订组织的推动以及 Intel、IBM、朗讯、Ericsson、Zynx 多家网络大公司的支持，ForCES 的技术结构成为目前国际上备受关注的实现开放可编程网络设计目标的体系结构。因此，转发与控制分离 (ForCES) 技术是实现开放架构网络的重要技术手段，IETF 在 2002 年专门成立 ForCES 工作组，开始有关 ForCES 技术和相关协议标准的研究制订工作。转发面由各类标准化的逻辑功能块 (Logical Functional Block, LFB)^[6]组成，并可由控制面按需要构造数据分组处理拓扑结构。转发面的编程性具体表现为模块间的拓扑构

造和模块的属性（Attributes）控制（如 configure/query/report）。典型的 LFB 如 IPv4/IPv6 Forwarder、Classifier、Scheduler 等。LFB 的格式由“FE 模型”（RFC5812）定义，而各种 LFB 的内容由“LFB 定义库”文件制订。控制面和转发面间的信息交换按照“ForCES 协议”（RFC 5810）实现。该体系能充分体现开放可编程网络的优点，即简洁的积木式开发以及不同控制面和转发面设备商间的可互操作性。

一个满足 ForCES 规范的网络件 ForCES 网络设备的基本结构如图 2-1 所示，RFC 3654（ForCES 需求分析）和 RFC 3746（ForCES 框架）对其进行基本定义。

如图 1 所示，一个满足 ForCES 标准的网络设备内有至少一个（或多个，用于冗余备份）控制件（Control Element，CE）和多达几百个转发件（Forwarding Element，FE）。CE 和 FE 间的通信通过称为“ForCES 协议”的标准协议完成，这个连接面称为 Fp 参考点（ForCES 控制接口），Fp 参考点可以经由一跳（Single Hop）或多跳（Multi-Hops）网络实现。2010 年 3 月，在经过 7 年多的努力后，IETF 完成了对 ForCES 协议的制订工作，成为 RFC 5810（ForCES 协议规范）。

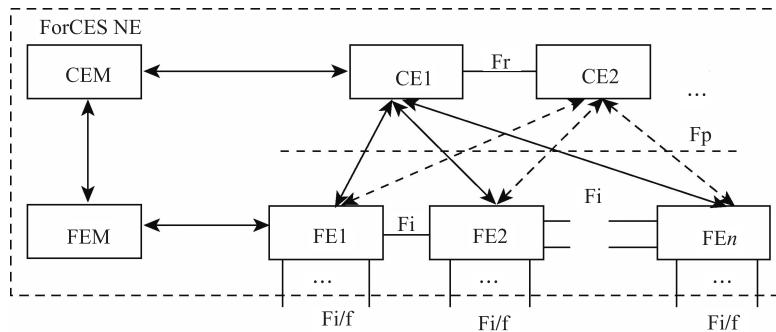


图 1 ForCES 网络件基本结构

ForCES 协议规定了 Fp 参考点上传递的两种消息的格式，这两种消息是控制消息和重定向消息。控制消息是包含 CE 对 FE 控制管理内容的消息，例如属性的配置和查询消息、能力和事件的上报消息。重定向消息是包含 CE 上所处理重定向数据分组的消息。从字面上理解，“重定向”数据分组不是指 FE 产生的数据分组，而是从外部到达 FE，需要由 FE “重新定向”到 CE 进行处理的数据分组；或者是 CE 产生的，需要经 FE “重新定向”到网络设备外部的数据分组。可能需要 CE 处理的数据分组主要有路由协议数据分组和网络管理数据分组等。

Fi/f 为各个 FE 对网络设备外的网络接口参考点，网络数据由此进出，并被该网络设备转发处理；Fi 为同一网络设备内各个 FE 间的相互连接接口协议，多个 FE 可以构成一个分布式的转发件网络，以完成复杂的转发功能。

Fr 为同一 ForCES 网络设备内各个 CE 间的连接协议。所有 CE 通过一个 CE 管理器（CE Manager，CEM）管理，所有 FE 通过 FE 管理器（FE Manager，FEM）管理，CEM 和 FEM 间也互相交换管理信息。但要注意的是，CEM 和 FEM 所做的只是一些基本的设置管理，如给各个 CE 和 FE 分配 ID 等，而对 FE 的全面管理是通过 CE 上面的软件

经由 ForCES 协议完成。CEM 和 FEM 可以被理解成 CE 或 FE 管理用的人机接口。图 1 所示为 CEM/FEM 实体在 CE/FE 外部，但是在物理上 CEM/FEM 很可能是嵌入 CE/FE 内部的。

ForCES 技术使得网络设备具有很强的模块化积木式特性。例如，主要是软件的 CE 和主要是硬件的 FE 可以在产品级分离，于是同一个网络设备内可以有不同厂商生产的 CE 和 FE。更进一步地说，在一个 FE 内，标准化的 LFB 也可以在产品级被分离、由不同厂商生产。同时，CE 具有灵活配置 FE 内各 LFB 的功能，通过构造不同的 LFB 拓扑结构，使网络设备能完成各种不同的服务业务。例如，当把网络设备从 IPv4 升级到 IPv6 时，只要通过 CE 加载相应的 IPv6 的 LFB 即可完成。

考虑到连接 CE-FE 链路的多样性和复杂性，传递 ForCES 协议消息的 ForCES 控制接口被进一步分为协议层（Protocol Layer, PL）和传输映射层（Transport Mapping Layer, TML），其结构如图 2 所示。这样做的目的是使 ForCES 协议的设计能独立于其所用的传输层。传输层可以是多样化的，如使用基于 SCTP、基于 TCP/UDP 甚至基于 ATM 网络的传输层等。

目前，ONF 提供的 SDN 技术白皮书^[7]中将 SDN 划分为 3 层，一分别是应用层（Application Layer）、控制层（Control Layer）和基础资源层（Infrastructure Layer），其实现实体都是网络节点，分别是应用节点、控制节点、基础资源节点。因此，浙江工商大学的 Forces 课题组提出了基于 ForCES 的 SDN 体系结构，如图 3 所示，其中考虑将 SDN 体系结构分为应用层、控制层和基础资源层。

图 4 所示为使用 ForCES 实现的 SDNFE 的系统框架，ForCES 中间件主要完成 ForCES 协议的交互过程。为了将 ForCES NE 改造为 SDNFE，关键的问题是需要解决 ForCES 网络设备中控制面和转发面两层资源到 SDNFE 间的映射问题，可采用重新定义 LFB 模型或者利用 ForCES 的 CE 直接控制 FE 中的 LFB。为此，可通过在 CE 中增加 SDNLFB 管理层和 ForCES + 协议中间件完成上述功能。

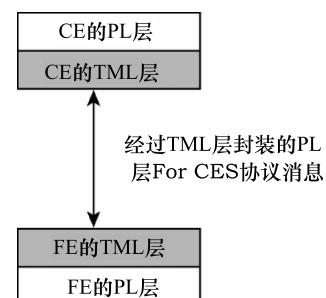


图 2 PL-TML 层分离结构

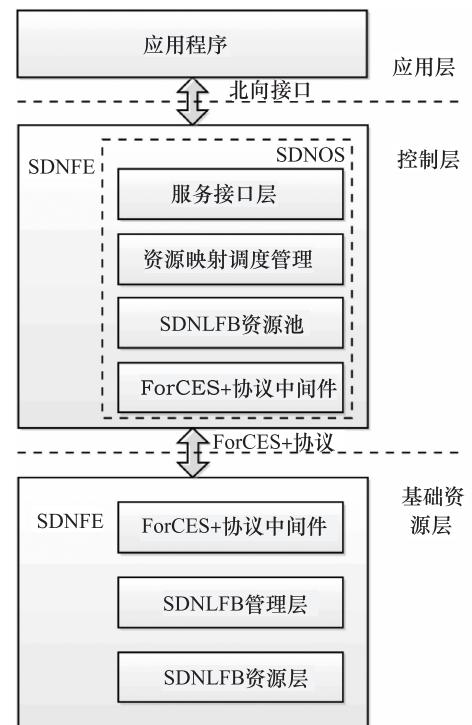


图 3 基于 ForCES 的 SDN 体系结构

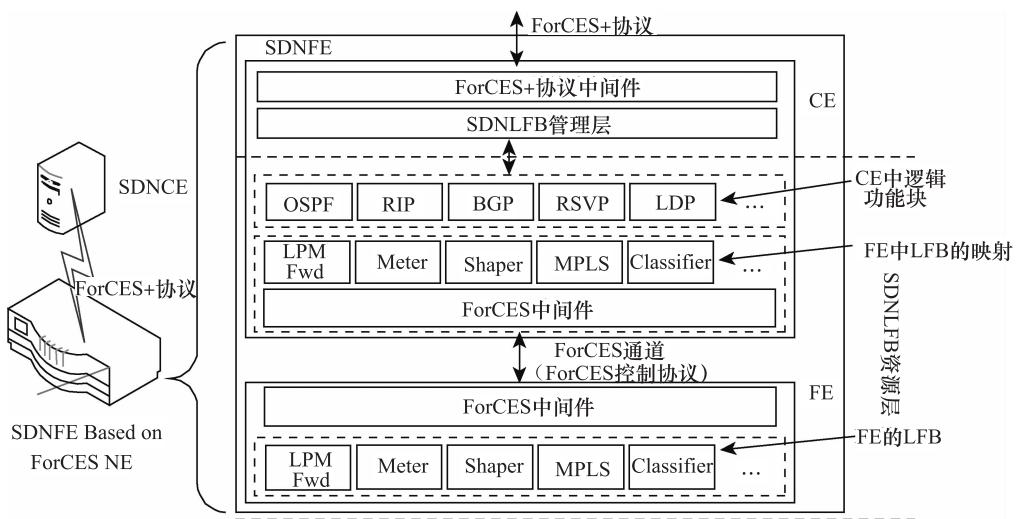


图 4 基于 ForCES 网络设备实现 SDNFE 的体系结构

2.1.2 SDN 体系结构

2.1.2.1 典型体系结构

SDN 体系结构如图 5 所示，即采用应用层（或称为业务层）、控制层（或称为服务层）和基础设施层（或称为转发层）的分离结构^[7]，以解决网络控制目标的多样性和路由的灵活性为出发点，引入可编程网络概念，从而解决控制面和数据面的僵化问题。

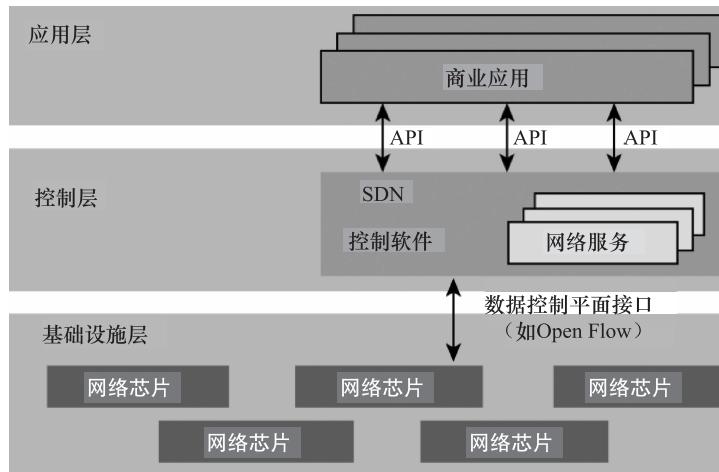


图 5 SDN 体系结构

解决决策面和数据面间直接控制的可伸缩性问题为后续体系研究的方向（即南北向接口）。2008 年斯坦福提出的 OpenFlow/SDN 技术范例（简称 OF/SDN），将体系结构分为控制面和数据面，分别由 OpenFlow 控制器和交换机构成，之间通过标准化 OpenFlow

(OF) 协议通信（如图 6 所示）。其中，控制器以集中方式控制由多个交换机所构成的域，其所属控制器通过一致性网络视图控制域间网络系统，控制器通过提供操作系统运行环境来支撑多种网络业务。

2.1.2.2 控制面技术

南北向接口的开放化增强了网络的创新能力，使网络运营商可以动态地配置、管理和优化底层的网络资源，实现灵活、可控的网络功能。然而，控制面仍面临许多技术挑战。

为了获得较好的响应能力和可伸缩性，一方面，单个 NOS 需要通过优化流水线结构和编程语言来改进交换机的控制吞吐量和所能控制的交换机数量；另一方面，由于网络自身的分布式特性和单个 NOS 的不可靠性，NOS 要能够通过分布式技术来减少交换机的时延和故障问题。以下分别阐述两种类型的 NOS。

(1) 集中式 NOS 技术

集中式 SDN 控制器研究工作以集中式控制为范例，以增强网络的可编程性为研究目的。先后有如下工作解决或改进控制器的可编程和可伸缩性：应用虚拟化技术并与云平台集成以及通过网络编程语言降低业务编程的复杂度。

(2) 分布式 NOS 技术

针对控制平面在可扩展性和通用性等方面的不足，Onix 提出了一整套面向大规模网络的分布式 SDN 部署方案，如图 7 所示，其体系统结构由网络控制逻辑、Onix、网络连接基础设施和物理网络基础设施 4 部分组成。

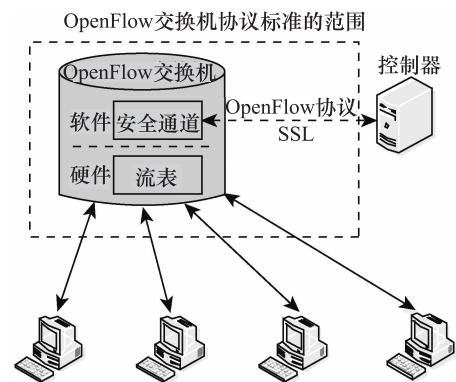


图 6 OpenFlow 体系结构及其基本组成

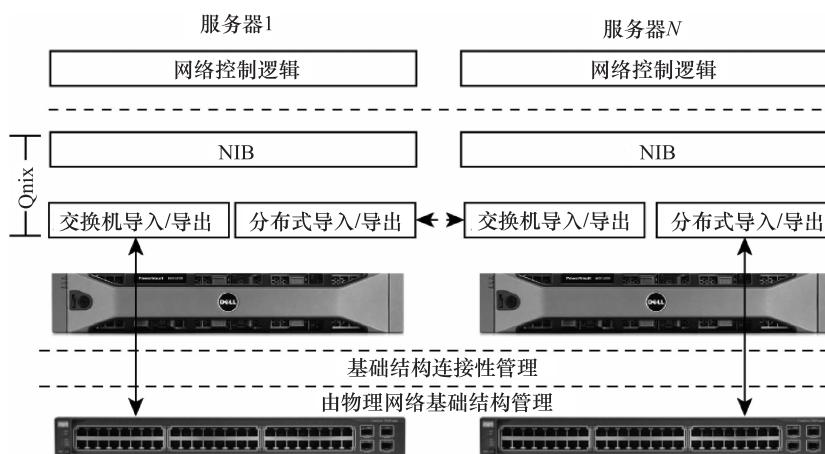


图 7 Onix 分布式网络体系结构及其组成部分

2.1.2.3 数据面技术

随着业务数量的增多以及对性能和可扩展需求的变化，高性能和可编程仍然是需要

解决的问题。在性能改进方面，主要是提高流的转发吞吐量，主要有如下两种方式。一种方式是对硬件体系结构及其相关算法进行改进。另一种方式是采用硬件加速方法辅助改进可编程交换的性能。

在可编程方面，现有 OF 流表从版本 1.0 ~ 1.3，将单级流表转变为多级流表，从而使得 IP 层以内的隧道等报文头部能够在流水线中得到匹配，增强了流表结构的表达性。然而，当引入新型网络协议之后，转发表自身也需要可扩展，数据面自身需具备演进能力，从而适应网络业务需求。目前主要有 3 种技术趋势：Google OpenFlow 2.0，该技术提出通过模块化以及上层软件抽象方式架构交换机，使得业务需求能够通过逻辑定义抽象来动态地对交换机的各个已有模块进行组合或定义新的模块功能，通过动态组织模块间的数据传输路径，实现所需的特定转发行为；交换机部分可重构，在该技术下，交换机运行于 FPGA 之上，在运行时交换机能够通过对 FPGA 上特定区域的逻辑进行动态编程来改变片上的转发逻辑，该方法需要上层提供复杂化的 FPGA 编程环境；交换机内的报文头部字段可重新编程，该技术允许用户定义新的协议头部字段模板，使得报文匹配时，可以根据模板来获取各个字段，进而匹配流表，但该方法将会干扰交换机的交换性能。

2.2 面向服务的新型网络体系结构

2.2.1 SOI

SOI (Service Oriented Internet)^[8]是由美国明尼苏达大学的 ChandrashekharJ 等提出的。顾名思义，就是采用面向服务的方式来描述未来互联网的结构。它是通过在现有网络层和传输层之间添加服务层（Service Layer）来建立一个面向服务的网络功能平台，属于演进式的研究思路，SOI 这种面向服务分发的网络设计思想具有灵活性强、统一性好、通用性优和可扩展的特点。它的基本体系结构如图 8 所示。

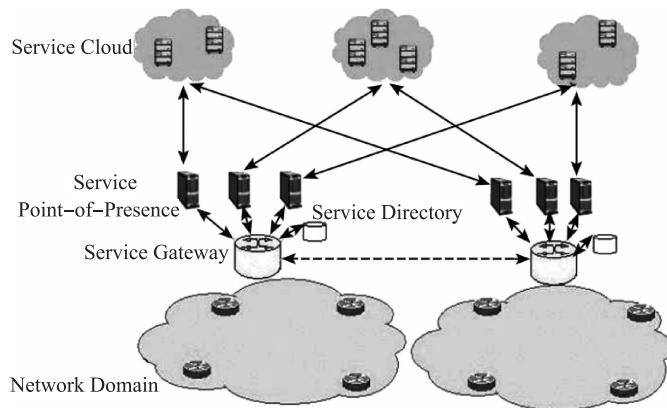


图 8 SOI 体系结构图

1) 在 SOI 体系结构中, 它将提供某类服务信息的各种服务实体 (Service Entity), 比如将内容服务器、代理服务器、缓存服务器、内容分发服务器等相关设备抽象到 SC (Service Cloud) 中, Service Cloud 可能是一类服务 (Service) 数据的来源, 也可能是转发一类服务相关数据的中间路由, 并且这个中间路由可能直接连接着被服务的用户。

2) 有了服务和数据对象的概念, 在网络传输中自然需要先对服务数据的来源和目的 SC、Object 等信息进行标记。SOI 对每个 SC 都采用长度固定 (32bit) 的 Service ID 来标记服务的来源 SC 和目标 SC, 这个标识由一个集中管理机构给定, 同 IP 地址的划分机制类似, 同时使用长度变化的 Object ID 来标记源 SC 中负责发送数据的源设备对象以及目的 SC 中实行最终数据分发的目标设备对象。Object ID 的长度之所以是变化的, 是因为其实现语法和语义是由 SC 内部提供的, 因而标记每个 Object 的 ID 长度也就是动态变化的, 这为 SC 的可扩展能力提供了基础, 能够有效地防止攻击, 保证了相应服务提供者的安全。图 9 是具体的服务数据的分组头格式。

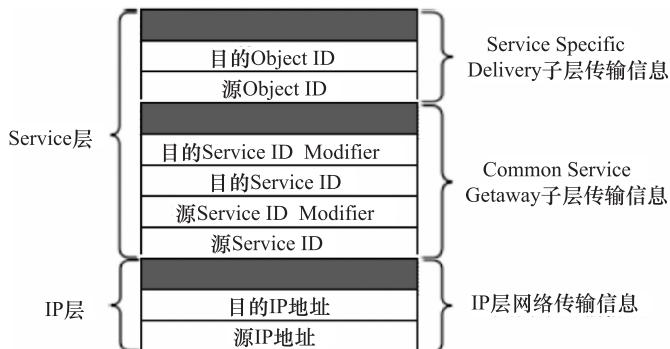


图 9 服务数据分组头格式

3) 由前面的介绍已知 SC 中的各种设备在现实中往往来自不同的 IP 网络域, 并且数据的传输还需要在网络域内部和域之间进行传输, 因此在转发服务数据的时候除了明确 SC 和 Object 信息, 还需要确认这些信息对应的具体网络域信息, 才能实现数据在 IP 网络中的传输。S-PoP (Service Point-of-Presences) 的作用正是处理从 SC 信息到具体 IP 网络域信息的映射, 实现的是 SC 与实际网络域之间的接口。对于域间移动的用户, S-PoP 还能提供动态更新 Object ID 与具体物理转发设备之间映射关系的功能, 满足了网络移动性的需求。

4) 虽然明确了源 SC、目的 SC 中对应的源 Object 和目的 Object 以及它们各自对应的 IP 网络域。但是服务数据尚缺少具体的传输起始路由器、具体的传输路径、中间需要经过的 IP 网络域、需要经过的路由等信息。这样的信息存储在 Service Gateway 中, Service Gateway 主要记录的是到达某个 SC 所需要经过的具体 IP 网络路由信息, 相关的信息则是 Service Gateway 通过 SGRP (Service Gateway Routing Protocol) 建立的。服务数据的层次结构以及数据在不同设备中传输时的分组头解析层次如图 10 所示。

5) SOI 建立了一个较为完整的面向服务的体系结构, 但依然存在相关的问题, 还需要进一步研究。比如, 服务和具体的 Service ID 如何建立映射关系, 服务数据的路由

转发需要包含哪些信息，如何确保数据传输的安全，数据实际传输的基本流程应该是怎样的，SGRP 如何建立起相关的路由信息等。这些基本问题都涉及整个 SOI 的可行性和对未来网络需求的满足程度，因此将针对上述问题按照逻辑排序之后，可进行进一步的解析。

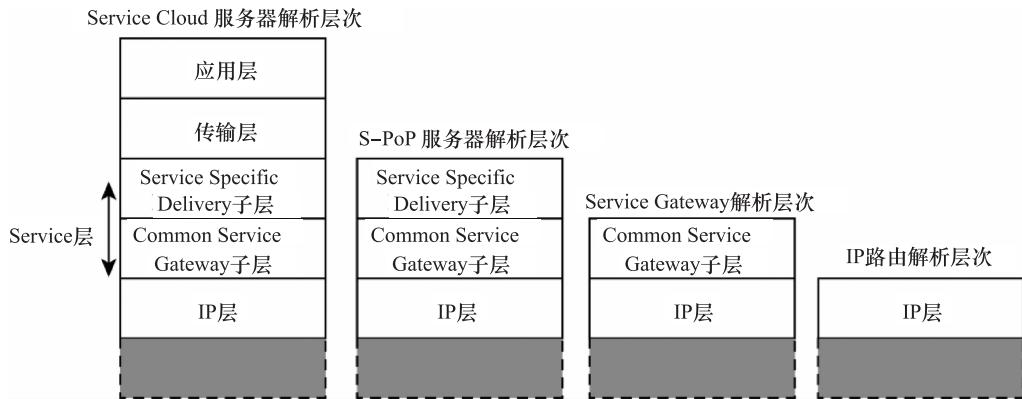


图 10 服务层和 SOI 协议栈

2.2.2 NetServ

NetServ^[9]是一个可编程的路由器体系结构，用于动态地部署网络服务。其设计的核心思想是服务模块化（Service Modularization）。NetServ 首先将网络路由节点中的可用功能和资源服务进行模块化，当需要在网络中建立一种相关的新服务的时候，NetServ 就会通过使用互联网络中的可用服务模块进行组合，最终形成相应的服务，构成服务的模块和多个模块构成的服务组件在 NetServ 中被统称为 Service Module。NetServ 中的服务模块是用 Java 中的 OSGi（Open Service Gateway Initiative）框架编写的，并通过发送 NSIS 信令消息实现部署管理。NetServ 还提供了虚拟服务框架（Virtual Services Framework），主要是为面向服务的网络体系结构中的路由节点提供相关安全保障、可控可管理、动态添加删除服务模块等功能。NetServ 需要使用的首要关键技术有两项：遥控模块路由器（Click Modular Router）^[10]和 Java 的 OSGi 框架。

图 11 是 NetServ 的体系结构以及在数据传输过程中获取服务的过程。这个体系结构中的阴影部分就是遥控模块路由器和 OSGi 框架的组成部分。Click Router 的相关 Element 是通过 C++ 类对 Element 进行描述和详细的功能定义，NetServ 则采用 Java 虚拟机中的 JNI 本地接口，实现对 Java 代码和 C++ 代码的相互调用；NetServ 中的 OSGi 启动器则有两个功能，第一是启动 OSGi 框架，将 Click 路由中的 Element 进行模块化，并形成 Building Block，第二是提供了一个 Java 类 PktConduit，PktConduit 使用 JNI 实现 OSGi 框架对 Click 路由器中 Element 服务的转换，从而形成路由器动态可扩展的功能模块集合。

NetServ 实现的第 3 个关键技术是节点的自我管理功能，网络中的相关服务功能模块

的注册、注销、更新、添加、删除、组合等功能的实现需要网络中多个路由器的参与，这就说明对于 NetServ 而言，管理控制模块显得尤为重要，并且相关的出错恢复、安全问题在自动管理功能中也显得尤为重要。总的来说，NetServ 的管理内容需要满足出错管理（Fault Management）、配置管理（Configuration Management）、计费管理（Accounting Management）、性能管理（Performance Management）、安全管理（Security Management）5 个需求（FCAPS）。整个 NetServ 节点的内部详细逻辑结构以及相关的信令分组和数据分组的处理过程参如图 12 所示。NetServ 实现的是自动的部署管理功能，而不是集中式的

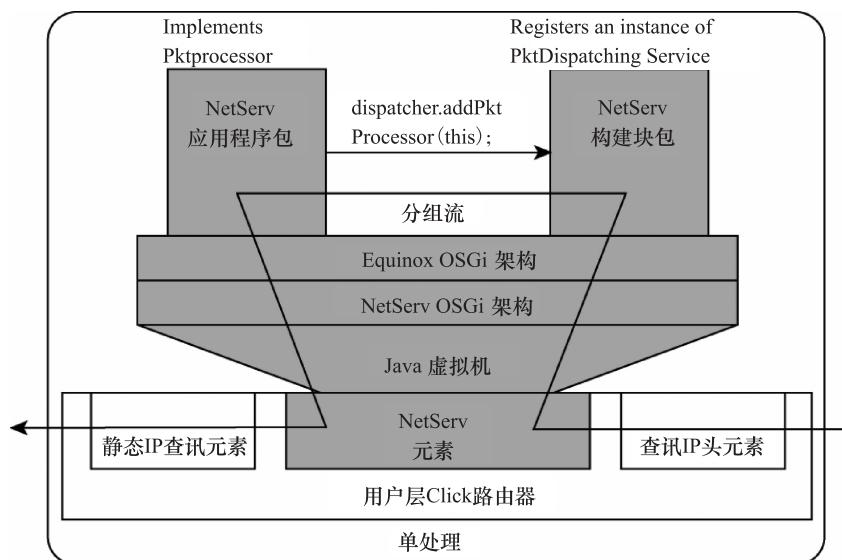


图 11 NetServ 原型系统体系结构

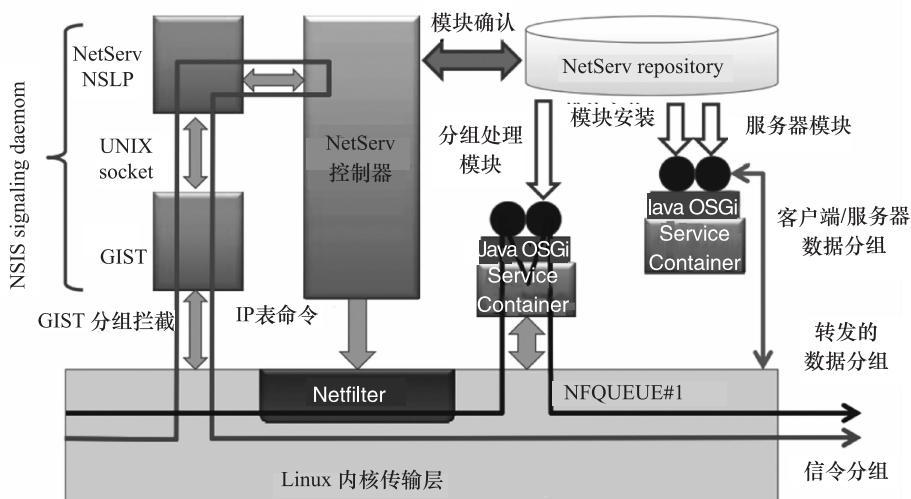


图 12 NetServ 节点内部体系结构

管理，这是由于服务本身在形成的时候，其网络边界和服务边界是不确定的。NetServ 依靠 NSIS 信令协议来实现，相关的信令能够用于 NetServ 节点的动态发现、内部服务模块的部署、NetServ 控制器则与信令进程、服务容器和节点的传输层等功能模块配合完成触发服务的动态添加/删除网络服务相关的功能模块。在网络中部署好的功能模块本身有自己的生命周期，需要在网络中通过信令确认自己的存在意义，否则网络服务中的各个功能模块会在超时之后直接被自动删除。最后，NetServ 控制器还有认证用户、建立/拆卸服务容器、提取或者分解功能模块等管理服务的策略。

2.2.3 COMBO

COMBO^[10]是欧盟 FP7 框架中关于网络体系结构的项目，主要研究固定和移动宽带接入/聚合网络收敛特性（Convergence of Fixed and Mobile Broadband Access/Aggregation Network），如图 13 所示。

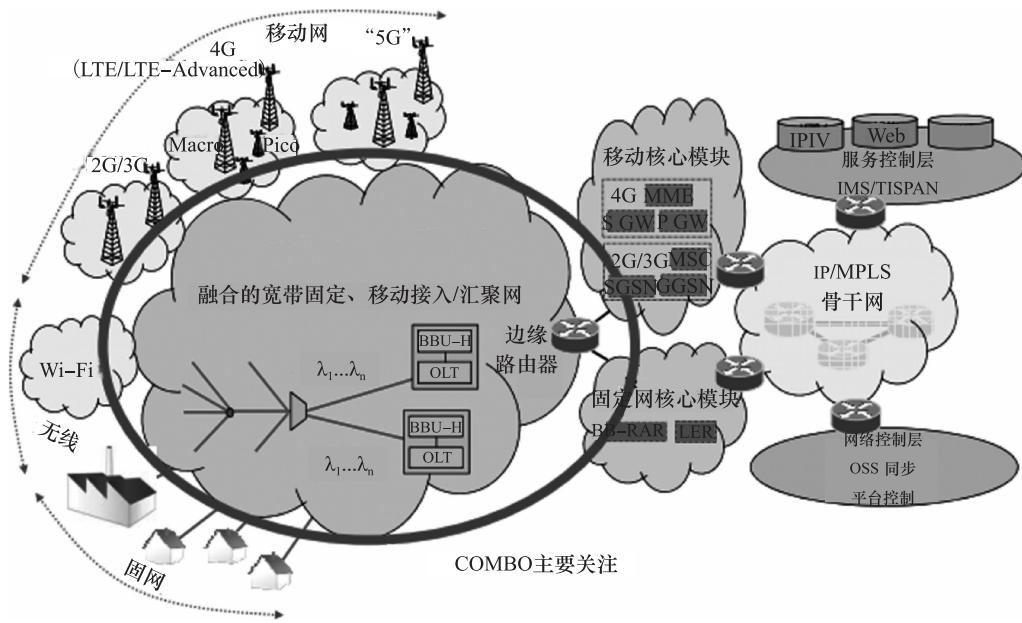


图 13 COMBO 关注于新的 FMC 宽带接入/汇聚网络体系结构的服务功能收敛性

COMBO 需要考虑网络体系结构中的收敛特性，它包括两个基本方面：一个是功能性的收敛，核心网络提供的服务在靠近边缘网络时会呈现发散的特点，比如移动网络中存在用户所接收的服务，与固网中的用户所接收的服务相同，如果能够得出这些靠近边缘网络的服务的收敛特性，那么 COMBO 就可以得到边缘网络与核心网络之间在网络服务上的差异性，从而为核心网络服务的分发提供策略依据，并对于得到更为长远的网络演进策略具有重大的理论意义；另一个是核心网络结构上的收敛特性，这对于网络资源的合理调度与配置以及网络的集中管理有指导意义，而对应着功能性的收敛，COMBO 能够更进一步地实现有效的资源分配策略和管理策略。下一代固定/移动融合的宽带接入点如图 14 所示。

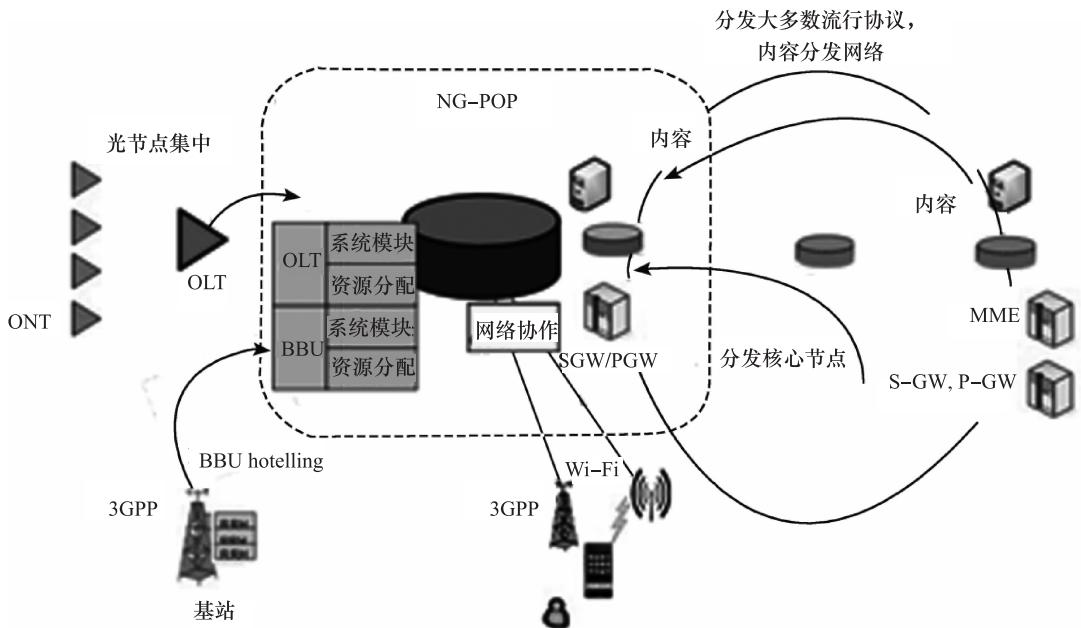


图 14 下一代固定/移动融合的宽带接入点

2.3 内容中心网络

2.3.1 NDN 体系结构

NDN (Named Data Networking)^[11]是由美国加州大学洛杉矶分校 Lixia Zhang 团队开展的研究项目，该项目由 FIA (NSF Future Internet Architecture) 资助，开始于 2010 年。NDN 的提出是为了改变当前互联网主机—主机通信范例，它使用数据名字而不是 IP 地址进行数据传递，让数据本身成为互联网体系结构中的核心要素。而由 PARC 的 Jacobson 在 2009 年提出的 CCN (Content-Centric Networking) 只是与 NDN 叫法不同，无本质上的区别。

NDN 中的通信是由数据消费者接收端驱动的。为了接收数据，消费者发出一个兴趣 (Interest) 分组，携带了和期望数据一致的名字。路由器记下这条请求进入的接口并通过查找它的转发信息库 (FIB) 转发这个兴趣分组。一旦兴趣分组到达一个拥有请求数据的节点，那么一个携带数据名字和内容的数据分组就被发回，同时发回的还有一个数据生产者的密钥信号。数据分组沿着兴趣分组创建的相反的路径返回到数据消费者。NDN 路由器会保留兴趣分组和数据分组一段时间。当从下游接收到多个要求相同数据的兴趣分组时，只有第一个兴趣分组被发送至上游数据源。在 NDN 中有两种分组类型：兴趣分组和数据分组。请求者发送名字标识的兴趣分组，收到请求的路由器记录请求来自的接口，查找 FIB 表转发兴趣分组。兴趣分组到达有请求资源的节点后，包含名字和内容以及发

布者签名的数据分组沿着兴趣分组的反向路径传送给请求者。通信过程中，兴趣分组和数据分组都不带任何主机或接口地址。兴趣分组是基于分组中的名字路由到数据提供者的，而数据分组是根据兴趣分组在每一跳建立的状态信息传递回来的，两者的格式如图 15 所示。

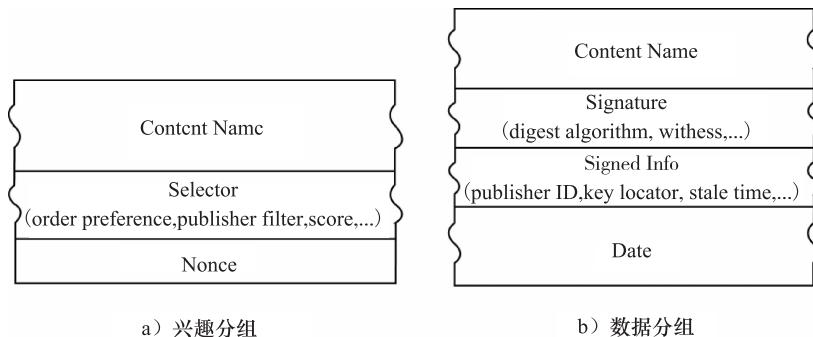


图 15 NDN 中的兴趣分组与数据分组格式

NDN 中引入了网内缓存的设计理念，与 CDN 代理服务器、P2P 缓存等边缘缓存相比，NDN 网内缓存在部署方式、缓存内容及获取方式等方面都有本质的不同，如表 1 所示。

表 1 缓存网络体系对比分析

网络体系	CDN	P2P	NDN
缓存部署目的	热点内容推进用户	缓解 P2P 带宽冲击	减少重复流量；缩短跳数
缓存部署者	内容提供商	网络提供商	网络提供商
缓存组织结构	层级式	扁平式	层级式 + 扁平式
针对业务类型	内容获取类业务	P2P 业务	所有业务
缓存内容粒度	文件	片断	片断
缓存内容来源	提供商向下推送	经节点转发的数据	经节点转发的数据
缓存部署位置	覆盖网方式	网络边缘路由器	所有路由节点内部
内容定位方式	DNS 重路由	通过协议分析	内容与名字的直接映射
内容获取方式	本地或上游服务器	P2P 缓存或者节点	逐跳式询问上游节点
内容更新方式	上游向下推送	节点替换策略	节点和路径缓存替换

NDN 体系包括命名系统、路由转发、缓存和 PIT 表等关键技术，其各自特点如下。

(1) 命名系统

命名系统是 NDN 体系结构中最重要的部分。NDN 采用分级结构的命名方式，例如，一个 PARC 产生的视频可能具有名字/parc/videos/WidgetA.mpg，其中“/”表示名字组成部分之间的边界（它并不是名字的一部分）。这种分级结构对代表数据块间关系的应用来说非常有用。例如，视频的版本 1 的第 3 段可能命名为/parc/videos/WidgetA.mpg/1/3。同时，分级允许大规模的路由。从理论上讲，利用扁平的内容名字实现全球路由转发是可能的。然而现实是：现有互联网的路由可扩展性严重依赖 IP 地址的分级结构。尽管全

局地检索数据要求全局的唯一性，但名字不需要全局唯一。专为局部通信的名字可能主要基于局部的内容，并仅要求局部路由（或局部广播）来找到对应请求的数据。

（2）路由和转发

NDN 基于名字的路由和转发解决了 IP 网络中地址空间耗尽、NAT 穿越、移动性和可扩展的地址管理 4 个问题。传统的路由协议，如 OSPF、IS-IS、BGP，也适用于基于名字前缀的 NDN 路由，NDN 路由器发布名字前缀公告，并通过路由协议在网络中传播，每个接收到公告的路由器建立自己的 FIB 表。NDN 节点的转发处理过程如图 16 所示，当有多个兴趣分组同时请求相同数据时，路由器只会转发收到的第一个兴趣分组，并将这些请求存储在 PIT 中。当数据分组传回时，路由器会在 PIT 中找到与之匹配的条目，并根据条目中显示的接口列表，分别向这些接口转发数据分组。NDN 节点的转发处理如图 16 所示。

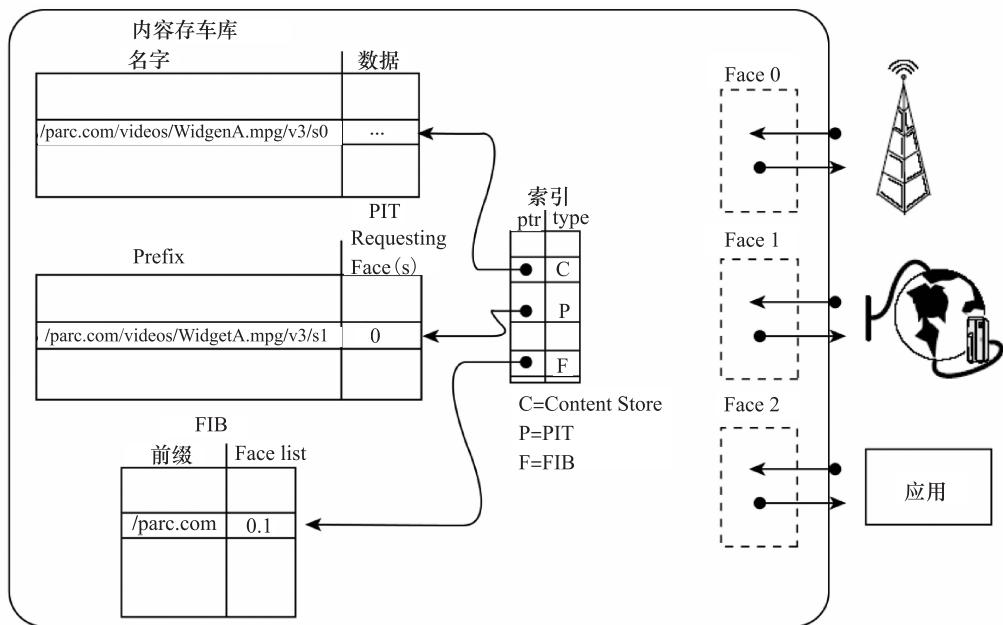


图 16 NDN 节点的转发处理

（3）缓存

一旦接收到一个兴趣分组，NDN 路由器首先检查内容库（Content Store），如果存在一个数据的名字在相应的兴趣分组下，则这个数据就会被作为响应发回。内容库的基本形式是现今路由器的缓存存储器。IP 路由器和 NDN 路由器都缓存数据分组，不同之处是，IP 路由器在转发数据之后不能再使用它们，而 NDN 路由器可以重用这些数据，以方便请求相同数据的用户。缓存在 NDN 中很重要，它可以帮助减少内容下载时延和网络带宽占用。NDN 采用 LRU 或 LFU 替换策略来最大限度地存储重要的信息。

（4）PIT

路由器将兴趣分组存放在 PIT（Pending Interest Table）中，该表中每个条目包含了兴

趣分组的名字和已经接收的匹配兴趣分组的接口集合。当数据分组到达时，路由器查找出与之匹配的 PIT 条目，并将此数据转发给该 PIT 条目对应的接口集合列表的所有接口，然后，路由器移除对应的 PIT 的条目，将数据分组缓存在内容库（Content Store）中。PIT 条目需要设置一个较短的超时时间，以最大化 PIT 的使用率。通常超时稍大于分组的回传时间。如果超时过早发生，数据分组将被丢弃。路由器中的 PIT 状态可以发挥许多关键作用：支持多播；限制数据分组的到达速率；控制 DDoS 攻击；实现 Pushback 机制等。

2.3.2 DONA 体系结构

DONA (Data-Oriented Network Architecture)^[12]是由美国加州大学伯克利分校 RAD 实验室提出的以信息为中心的网络体系结构。DONA 对网络命名系统和名字解析机制做了重新设计，替代现有的 DNS，使用扁平结构、Self-Certifying 名字来命名网络中的实体，依靠解析处理器（Resolution Handler）来完成名字的解析，解析过程通过 FIND 和 REGISTER 两类任播原语实现。

DONA 的命名系统是围绕当事者进行组织的。每个当事者拥有一对公开—私有密钥，且每个数据、服务或其他命名的实体（主机、域等）和一个当事者相关联。名字的形式是 P: L，P 是当事者的公开密钥的加密散列，L 是由当事者选择的一个标签，当事者确保这些名字的唯一性。当一个用户用名字 P: L 请求一块数据并收到三元组 <数据，公开密钥，标签> 时，他可以通过检查公开密钥的散列 P 直接验证数据是否确实来自当事者，且标签也是由这个密钥产生。

DONA 名字解析使用名字路由的范式。DONA 的名字解析通过使用两个基本原语来实现：FIND (P: L) 和 REGISTER (P: L)。一个用户发出一个 FIND (P: L) 分组来定位命名为 P: L 的对象，且名字解析机制把这个请求路由到一个最近的复制，而 REGISTER 消息建立名字解析的有效路由所必须的状态。每个域或管理实体都将有一个逻辑 RH，当处理 REGISTER 和 FIND 时，RH 使用本地策略。每个用户通过一些本地配置知道他自己本地 RH 的位置。被授权用名字 P: L 向一个数据或服务提供服务的任何机器向它本地的 RH 发送一个 REGISTER (P: L) 命令，如果主机向当事者关联的所有数据提供服务（或将进入的 FIND 分组转发给一个本地副本），注册将采用 REGISTER (P: *) 的形式。每个 RH 维护一个注册表（Registration Table），将名字映射到下一跳 RH 和复制的距离（也就是 RH 的跳数或一些其他向量）。除了各种 P: L 的单个条目外，P: * 有一个单独的条目。RH 采用最长前缀匹配法，如果一个 P: L 的 FIND 请求到达，且有一个 P: * 的条目而没有 P: L 的条目，RH 会使用 P: * 的条目；当 P: * 的条目和 P: L 的条目都存在时，RH 将会使用 P: L 的条目。当一个 FIND (P: L) 到达时的转发规则是：如果注册表中存在一个条目，FIND 将被发送到下一跳 RH（如果有多个条目，则根据本地策略选择一个最接近的条目）；否则，如果 RH 是多宿主的，RH 将把 FIND 转发到它的双亲（如它的供应者），使用它的本地策略来选择，其过程如图 17 所示。

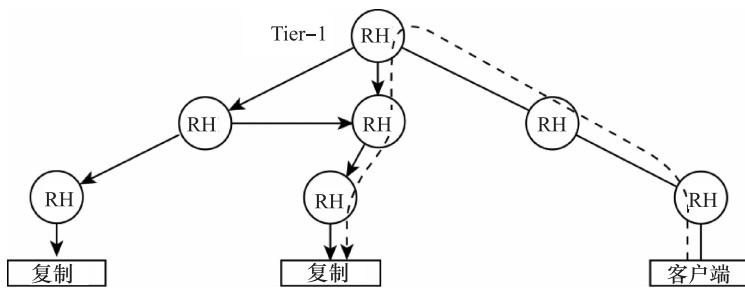


图 17 DONA 名字路由过程示例

FIND 分组的格式如图 18 所示。DONA 在 IP 头部和传输层头部之间插入一个填隙片。DONA 提供的基于名字的路由确保数据分组到达一个合适的目的地。如果 FIND 请求到达一个 1 级 AS 且没有找到有关当事者的记录，那么 1 级 RH 会返回一个错误消息给 FIND 信息源。如果 FIND 没有定位一个记录，对应的服务器会返回一个标准传输级响应，为了实现这个目标，传输层协议应该绑定到名字而不是地址上，但是其他方面不需要改变。同样地，当请求传输时，应用协议需要修改为使用名字而不是地址。事实上，当在 DONA 上实现时，许多应用会变得简单。例如 HTTP，注意到 HTTP 初始化中唯一关键的信息是 URL 和头部信息；考虑到数据已经在低层命名，不再需要 URL，同时，如果数据的每个变量给定一个单独的名字，那么头部信息页将变得多余。接收到 FIND 后发生的数据分组的交换不是由 RH 处理的，而是通过标准 IP 路由和转发被路由到合适的目的地。在这种意义上，DONA 并不要求修改 IP 基础结构。

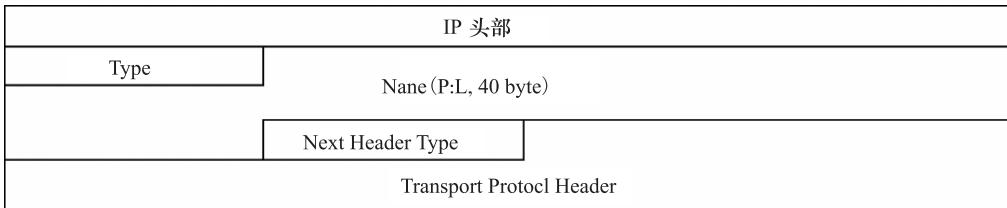


图 18 FIND 分组的协议头部

2.3.3 PURSUIT 体系结构

PURSUIT (Publish Subscribe Internet Technology, 曾为 Publish Subscribe Internet Routing Paradigm, PSIRP)^[18]是由欧盟 FP7 资助的研究项目，于 2008 年 1 月启动。PURSUIT 旨在建立一个以信息为中心，基于“发布”—“订阅”模式的通信架构来替换传统的端到端模式。发布者只负责发布信息，用于响应来自请求者的订阅服务。为了取代目前的 IP 协议，PURSUIT 提出了一个完整的“发布”—“订阅”协议栈，实现对现有网络的彻底改造。

(1) 命名系统

PURSUIT 采用唯一的标识对来命名数据对象。标识对包含了作用域标识 (Scope ID，

SID) 和匹配标识 (Rendezvous ID, RID)，例如 < SID, RID >。类似于 DONA，匹配标识采用扁平的命名机制，而层次化的作用域标识决定了该数据对象的被访问权限、认证信息、可达性信息和可用性信息等。例如，一个发布者将一张照片同时发布在家庭域和工作域中。在不同的域中，对应的作用域标识不同，对应的被访问权限也不同。尽管如此，该照片却有一个唯一的匹配标识。

(2) 名字解析和内容转发

PURSUIT 网络体系结构由三种功能实体组成：匹配 (Rendezvous)、拓扑管理 (Topology Management) 和转发 (Forwarding)，如图 19 所示。多个匹配节点 (Rendezvous Nodes, RN) 构成分布式哈希表 (DHT) 来实现匹配功能，为订阅消息找到网络中的发布者。具体来说，当发布者向网络发布一个数据对象 (SID, RID) 时，他会向本地的 RN 发送一个发布消息 (PUBLISHmessage)，而本地的 RN 将发布消息发送至与数据对象的 SID 匹配的 RN 处。当订阅者订阅该数据对象时，会将订阅消息 (SUBSCRIBEmessage) 发送至本地的 RN。该 RN 会根据数据对象的 SID 将订阅消息发送至与 SID 匹配的 RN 处。

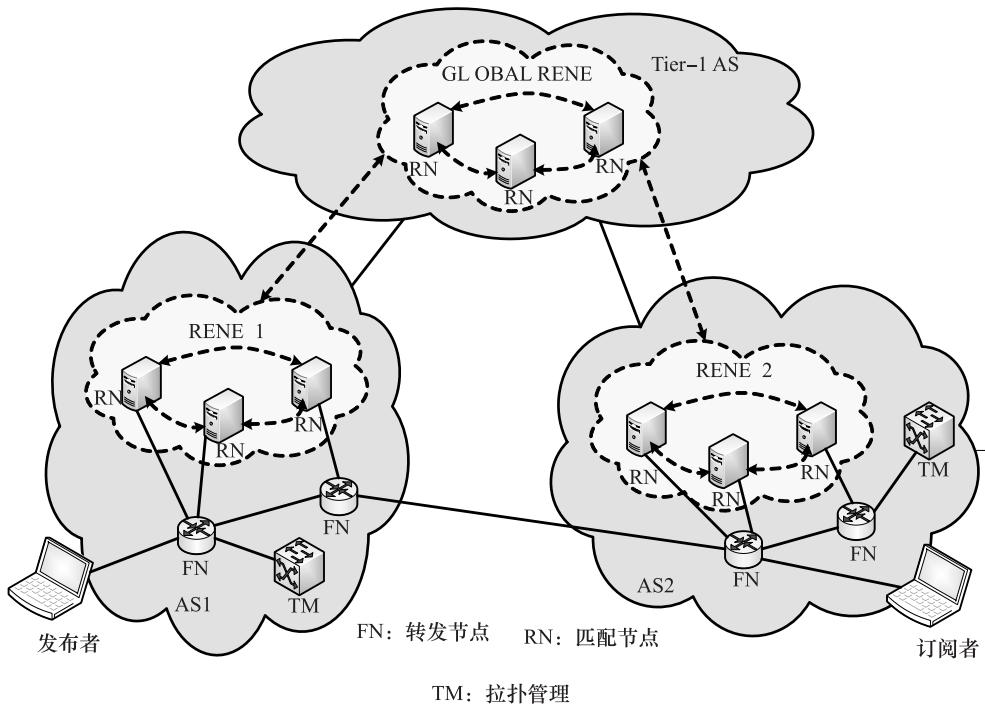


图 19 PURSUIT 体系结构

根据发布消息和订阅消息中数据对象的 RID，RN 能够匹配订阅消息和发布消息，同时将发布者和订阅者的信息通告给拓扑管理功能实体 (Topology Management, TM)。TM 在发布者和订阅者之间找到一条合适的路径，用于传递数据。同时，TM 将路径信息通过开始发布消息 (START PUBLISHmessage) 通告给发布者。最后由转发节点 (Forwarding Nodes, FN) 将数据对象传递到订阅者处，完成数据对象获取。因为 PURSUIT 中名字解析和内容路

由是完全沿着不同的路径，所以 PURSUIT 有效地实现了名字解析和内容路由分离。

(3) 缓存

对于缓存，PURSUIT 要求转发节点（FN）能够在转发数据包的同时，在本地保留一份数据包的副本。所以，PURSUIT 将数据对象沿其转发路径缓存（也就是 on-path caching）。同时，由于名字解析和内容路由分离，PURSUIT 也能够支持将数据对象的副本保存在网络中其他的缓存位置（例如缓存空间富余的内容路由器），用于提高缓存资源的利用率。

(4) 移动性

多播和缓存能够有效促进 PURSUIT 对移动性的支持。当用户开始移动时，发布者可以使用多播技术向用户潜在的移动位置发送内容。当切换完成时，用户可以通过临近的缓存获取数据对象。而当内容源移动时，需要内容源向拓扑管理功能实体 TM 通告新的位置信息。

2.3.4 SAIL 体系结构

2010 年 8 月，欧盟建立了 SAIL (Scalable and Adaptive Internet Solutions，曾为 Architecture and design for the future Internet: 4WARD))^[19]项目，目的在于演进当今的网络体系结构，以满足未来互联网的需求。SAIL 项目由瑞典爱立信公司主持，并受 FP7 计划资助，于 2013 年 2 月结束。SAIL 结合了 NDN 和 PURSUIT 中的元素，形成了混合式的体系结构，如图 20 所示。

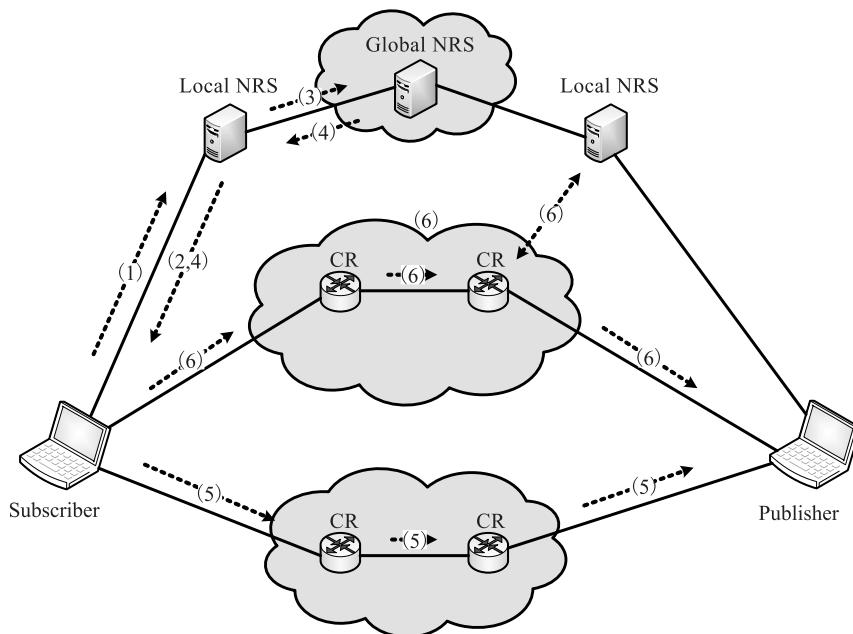


图 20 SAIL 体系结构

(1) 命名系统

SAIL 采用“略微扁平”的数据对象命名机制。具体来说，SAIL 并没有一个严格的命名规则，可以采取扁平的命名机制，也可以采取不携带任何位置或者拓扑结构信息的层次化命名。所以，SAIL 的内容名字可以是哈希值，也可以是一串字符串。但是 SAIL 要求将内容名字划分为两个部分，例如 A/L。A 部分的名字可用于全网，并在全网范围内唯一，而 L 部分只在本地有效。

(2) 名字解析

SAIL 支持三种名字解析和内容获取方式，一种是基于名字解析服务，类似于 DONA；另一种是名字解析和数据转发绑定，类似 NDN；还有一种是结合上述两者的混合解析方式。当采用名字解析服务时，SAIL 提出分布式的名称解析系统。其中本地 NRS 负责解析内容名中的 L 部分，如图中第 1 步所示。如果有相对应的条目，则 NRS 返回对应的位置信息（如第 2 步所示）；如果没有，NRS 则将解析请求发送至全局 NRS（如图中第 3 步所示）。全局 NRS 负责解析 A 部分，并返回对应的位置信息，如图中第 4 步所示。

当 SAIL 将名字解析和数据转发绑定时，内容源需要使用路由协议（类似 OSPF 或者 BGP 协议）通告内容的可达性信息。内容路由器保存该可达性信息，并依据可达性信息逐跳转发来自客户端的服务请求，如图中第 5 步所示。

混合解析模式结合了上述的两种解析方式。当服务请求被内容路由器逐跳转发时，如果路径上的某个内容路由器找不到对应的可达性信息，则将服务请求发送至本地 NRS 和全局 NRS，根据获得的内容源位置来获取数据对象。混合解析模式如图中第 6 步所示。

(3) 缓存

SAIL 提出了分层的缓存机制。位于接入域的内容路由器作为缓存机制的底层，而其他内容路由器作为上层。上层的内容路由器拥有更多的缓存空间，以存储更多的数据对象。

(4) 移动性

当主机移动时，主机需要将新的位置信息通告给 NRS，同时通告给每一个正在跟移动主机通信的对端。当内容源移动时，如果内容源并没有移动出当前的区域，则只需要向本地 NRS 通告新的位置信息，而不用向全局 NRS 进行通告。

2.3.5 COMET 体系结构

COMET (Content Mediator Architecture for Content-aware Networks)^[20] 是欧盟 FP7 项目的一部分。COMET 根据用户的传输需求、服务偏好以及当前的网络状态，为用户的服务请求选择合适的信息源。COMET 体系结构中的核心组件为 CMP (Content Mediation Plane)，它位于网络提供者与资源服务器之间，既保存了网络当中的服务资源信息，又能感知到网络底层的设备状态。与其他 ICN 体系结构不同的是，COMET 允许服务的提供者和订阅者指定明确的服务资源位置信息偏好。例如，某个服务订阅者可以只请求一个指定国家的图书资源；某个服务提供者可以只将自己的服务资源提供

给特定国家的用户。

COMET 体系结构中的通信流程如图 21 所示，服务提供者（Publisher）发送 REGISTER 消息给本地的 CRS（Content Resolution System），CRS 收到该注册消息后为其分配名字并且存储其位置信息；该信息通过 PUBLISH 消息逐级向上层 AS 的 CRS 汇报。当用户请求该资源时，向其本地 CRS 发送 CONSUME 消息；如果本地 CRS 未存储该资源的位置信息，则向其上层 CRS 查询，直到找到该资源为止。服务提供者可以指定该请求消息传播的区域，从而限制其获取资源的范围。当查找到服务提供者的位置信息后，CONSUME 消息会根据 CRS 指向的位置被转发至服务提供者处。在上述过程当中，每一个中间自治域的 CRS 会向 CaR（Content-aware Routers）配置相应的转发规则，这些转发规则从服务提供者延伸至服务请求者。最后，服务提供者利用这些转发规则将服务数据传送给请求者，完成一次完整的服务请求和应答过程。

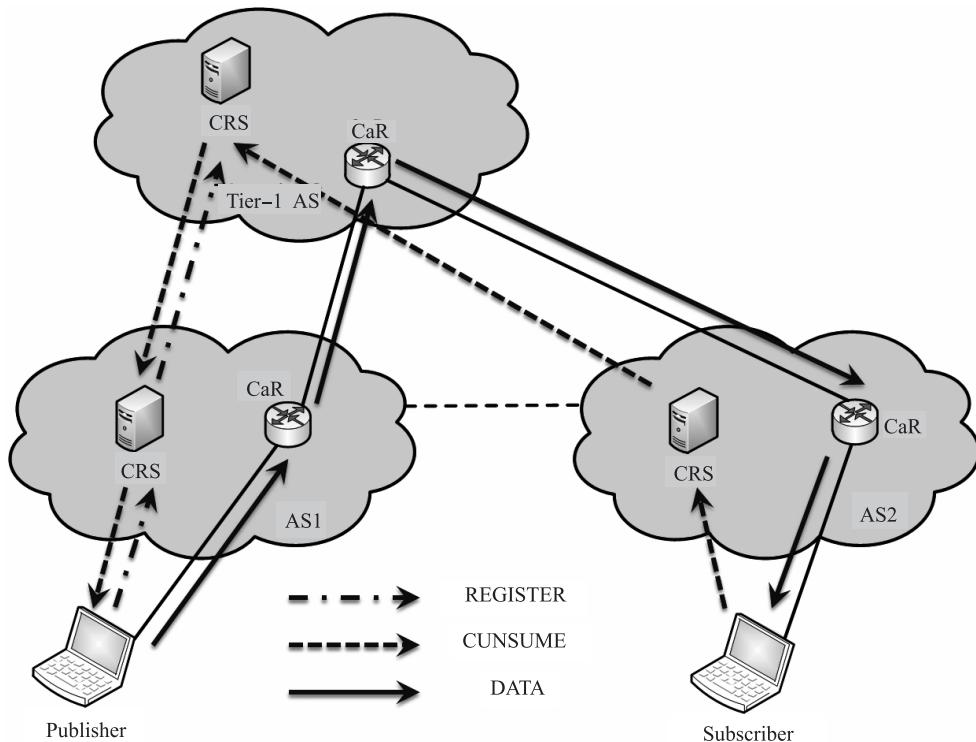


图 21 COMET 体系结构

COMET 体系结构特点如下：

(1) 缓存

COMET 同时支持 on-path 和 off-path 的缓存。COMET 已经提出了两种 on-path 的缓存机制：ProbCache 缓存机制和节点中心度缓存机制。ProbCache 是一种基于概率的缓存机制，路由器根据其与内容提供者和请求者之间的距离，利用特定的算法来计算缓存该内容的概率。节点中心度指的是该节点位于拓扑中所有节点对之间最短路径上的次数，节

点中心度缓存机制仅仅将服务资源缓存在节点中心度最高的路由器上，因为位于很多最短路径上的缓存内容更容易被命中。计算 CaR 的节点中心度要求每个 CaR 都知道全局的拓扑信息，因此，COMET 设计了一种简化的算法，使路由器只需要知道邻居的链路信息即可。仿真结果表明，COMET 的两种缓存算法均比 NDN 的缓存算法命中率高、可扩展性强。

(2) 移动性

COMET 利用位于边缘网络的 CaR 来为用户的移动性提供支持。这些路由器可以追踪用户的移动性信息并且对未来用户的位置做出预测。当服务请求者从一个 CaR 移动至同一个域的另外一个 CaR 时，后者能够主动从先前的 CaR 处获取该用户的环境信息。

(3) 安全性

COMET 继承了其他 ICN 体系结构在安全性方面的特点，安全策略的使用要根据具体的命名方式来决定。例如，当为了方便名字的聚合而使用连续名字来命名服务时，服务的命名就不能使用像 DONA 这样自证明的方式来保证安全性。另外，在 COMET 中采用 AS 级的路径来进行路由而不是利用全球范围的地址，可以有效防止攻击者发起不可探测的网络攻击。

2.3.6 CONVERGENCE 体系结构

CONVERGENCE^[21]项目是欧盟 P7 项目的一部分。该项目一个突出的特点就是试图利用 IP 网络中现有的功能来完成网络的平滑过渡。例如，CONVERGENCE 体系结构中制定了规则将大的数据分为若干个载体数据包（比如 IP 数据包）。另外，CONVERGENCE 体系结构定义了一个 IP 数据包头可选字段，用来承载 CONVERGENCE 消息包头中的关键信息，从而使路由器能够区分对待含有 CONVERGENCE 消息的 IP 数据包。在 CONVERGENCE 体系结构中，服务名字包含两个部分：名字空间标识（namespace ID）和名字（name）。CONVERGENCE 中默认的命名方式类似于 DONA 中的 P: L 对，但也可以用类似于 NDN 中的层次化命名方式来命名服务。

CONVERGENCE 中的通信流程如图 22 所示。服务请求者向网络发送一个 INTEREST 消息来请求某一服务，该消息被 BN（Border Nodes）逐跳转发至服务提供者或者存有数据副本的 IN（Internal Nodes）。服务提供者则将服务数据沿原路径发送至服务请求者处。在 CONVERGENCE 中，BN 并不保存每个名字的路由信息，而是仅仅保留其中的一部分。当 BN 无法获得某个 INTEREST 消息的路由条目时，则询问本地的 NRS（Name Resolution System）服务器。当 INTEREST 消息被路由时，会记录其沿途经过的 BN 的地址信息，因此服务提供者可以根据这些信息将数据回传给服务请求者，而不用在路由器中维护任何转发状态。另外，相邻的 BN 之间不一定是直连的链路，可以是由 IP 路由器组成的多跳链路。

在 CONVERGENCE 体系结构中，名字的解析和数据的路由是耦合的，因为 DATA 消息会沿着 INTEREST 消息的转发路径返回。

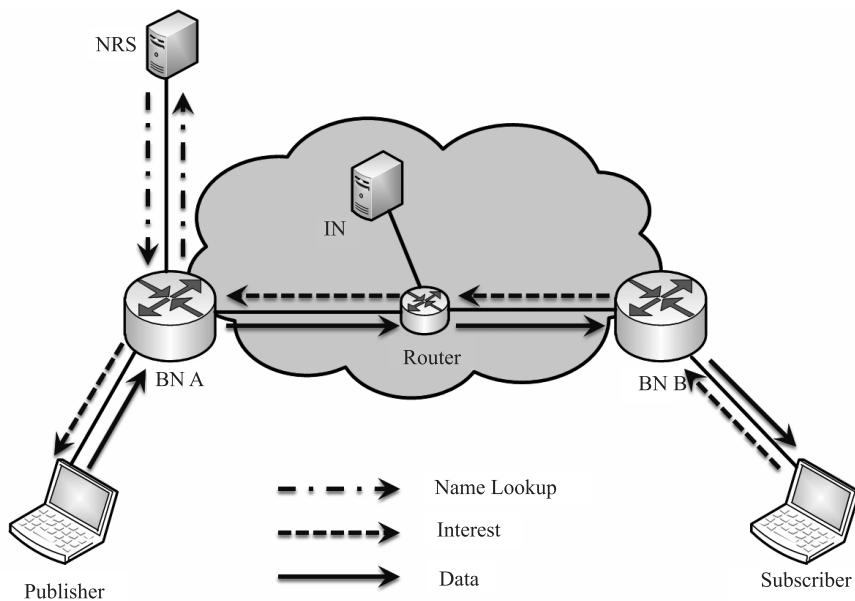


图 22 CONVERGENCE 体系结构

CONVERGENCE 体系结构特点如下：

(1) 缓存

CONVERGENCE 体系结构支持类似于 NDN 中的 on-path 缓存机制。off-path 的缓存可以通过 IN 中的服务副本向 NRS 注册来实现，但是这种缓存的开销是未知的，因为 NRS 的通信机制还未被定义。

(2) 移动性

与 NDN 相同，CONVERGENCE 体系结构支持服务请求者的移动，只需服务请求者向网络中重传 INTEREST 消息即可。但是，CONVERGENCE 中请求者移动切换的性能需要考虑，因为其并不在转发节点中保存路由状态，因此重传的 INTEREST 消息必须被发送至服务提供者，这可能会造成移动切换时延的增加。另外，内容源移动性的支持需要向 NRS 更新路由信息，而这种更新的开销目前也是未知的。

(3) 安全性

CONVERGENCE 体系结构继承了 NDN 体系结构中基于内容的安全机制，即每一个 DATA 消息包含一个数字签名。鉴于签名认证的巨大开销，DATA 消息的大小要远大于承载数据包的大小。另外，CONVERGENCE 体系结构中只在服务请求者端进行 DATA 消息级的安全认证，并不在 BN 中进行承载数据包级的安全认证。

2.4 面向移动性的新型网络体系结构

2.4.1 MobilityFirst

MobilityFirst^[13]项目是 NSF 未来网络体系结构（Future Internet Architecture）项目的

一部分，目标在于为移动服务开发高效和可伸缩的体系结构。MobilityFirst 项目面向移动平台和应用，假定互联网中移动终端（如手机、ipad 等）的数量大大超过固定终端的数量。这种假设提供了独特的机会来设计一种基于移动设备和应用的下一代互联网。

2.4.1.1 MobilityFirst 体系结构

MobilityFirst 体系结构的设计目标是：用户和设备的无缝移动；网络的移动性；对带宽变化和连接中断的容忍；对多播、多宿主和多路径的支持；安全性和隐私；可用性和可管理性。这些需求由以下协议成分实现。

(1) 身份和网络地址明确分离

MobilityFirst 明确地把可读的名字、全局唯一的标识符（Globally Unique Identifier, GUID）和网络位置信息区分开来。名字认证服务（Name Certification Service, NCS）安全地把可读的名字与全局唯一标识符绑定起来，而全局名字解析服务（Globally Name Resolution Service）把 GUID 映射到网络地址（Network Address, NA）。通过使 GUID 成为密码的可认证标识符，MobilityFirst 提高了可信性。相反的，通过明确的分离网络位置信息和 GUID，MobilityFirst 可以确保无缝移动。

(2) 分散的名字认证服务

不同的、独立的 NCS 机构能够验证名字和相应 GUID 之间的绑定，由于不同机构有可能对名字对应的 GUID 有争议，因此端用户可以选择一个值得信任的 NCS，使用基于法定人数的技术来解决 NCS 上的争议。

(3) 大规模可伸缩的全局名字解析服务

GNRS 是 MobilityFirst 的最核心的成分之一，它能够支持大规模的无缝移动。这里的大规模，指的是 100 亿台移动设备每天移动 100 个网络，相当于更新开销为大约 1 000 万/s。与此对应，DNS 过度依赖于高速缓存并需要多天来更新一个记录。因此，设计一种大规模的、可伸缩的、分布式的 GNRS 是 MobilityFirst 的一个重要挑战。

(4) 广义的存储感知路由

MobilityFirst 利用路由器中网内的存储来解决无线接入网络带宽变化和偶然的连接中断。CNF 体系结构的早期工作论证了存储以及存储感知路由的好处，存储感知路由算法在做转发决定的时候还需考虑长期和短期的路径质量。全局存储感知的路由（Generalized Storage-Aware Routing, GSTAR）协议在 CNF 存储中融入了时延容忍能力来为无线接入网络提供无缝的解决方案。

(5) 内容和上下文感知服务

MobilityFirst 中的网络层被设计成内容感知的，即它主动地帮助内容检索，而不像在现有网络中，提供一个原语把数据分组发往目的地。MobilityFirst 通过给内容分配密码可认证的 GUID 来达到这个目标。MobilityFirst 还把基本的设备和内容 GUID 扩展到更灵活的设备或用户组，如一个公园中所有的移动设备。

(6) 计算和存储层

现在的互联网急需可发展性。为此，MobilityFirst 路由器支持计算和存储层快速引进新的服务，从而使其对现有用户性能的影响降到最小。

2.4.1.2 协议设计

MobilityFirst 协议体系结构基于网络对象名字和网络地址的分离，基于特定应用的名字认证服务可以把可读的名字翻译成一系列网络地址。NCS 把可读名字翻译成唯一的 GUID，GUID 可被用作网络对象（如设备、内容、传感器等）的权威性标识符，同时，GUID 也是一个公钥，可以提供一种机制来验证和管理所有网络设备和对象的可信性。这个框架也支持基于上下文的描述符概念，如图 23 所示，可以通过上下文名字服务把杭州的所有出租车解析成一个特别的 GUID，把所有杭州的出租车作为一个动态的多播组。一旦将一个 GUID 分配给一个网络对象，就在 GUID 和网络地址之间建立了一个映射。GNRS 通过提供移动设备的当前接入点来支持动态的移动性。

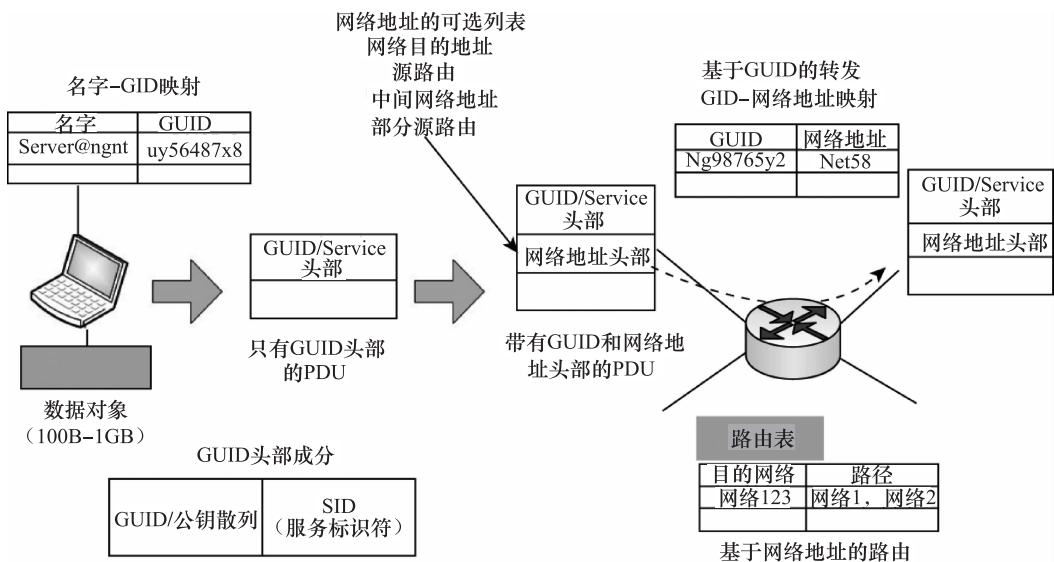


图 23 MobilityFirst 中混合的 GUID/NA 数据分组头部

Mobility 体系结构的另一个特点就是路由器中存在网内的存储，这可以使用存储感知的路由协议，把 PDU 暂时存储在路由器中而不是转发到目的地，以处理质量不好的链路和连接中断。用一个可信的逐跳传送的协议在路由器之间传送数据分组而不使用 TCP/IP 中端到端的方法。

MobilityFirst 协议栈的另一个重要特点是服务灵活性，它具有多播、任播、多路径和可以作为路由协议中完整功能的多宿主模式。这些服务是基于移动应用，常常是基于上下文而提出的。

GUID 机制考虑多播或任播到 GUID 相关的一系列网络地址的上下文和内容的寻址能力。一个比较有趣的难以用传统的 IP 处理的例子是“双归属主机”，一个用户的笔记本电脑有两个或多个无线接口连接到不同的网络，服务目标是发送 PDU 到至少一个接口。

2.4.2 HIP

HIP (Host Identity Protocol)^[14]在传统的 TCP/IP 体系网络中引入了一个全新的命名

空间——节点标识 (HI)，在传输层和网络层之间加入了节点标识层 (Host Identity Layer)，用于标识连接终端，安全性和可移动性是其设计中尤为推崇的特性。HIP 的主要目标是解决移动节点和多宿主问题，保护 TCP、UDP 等更高层的协议不受 DoS 和 MitM 攻击的威胁。

节点标识，实质上是一对公私钥对中的公钥，节点标识空间基于非对称密钥对。由不同公钥算法生成不同长度的 HI，HIP 再将 HI 进行散列来得到固定长度 (128 bit)、固定格式的 HIT，以便作为 HIP 报文的节点标识字段 (可包含在 IPv6 扩展头内)。为了兼容 IPv4 地址协议和应用程序，HIP 还定义了局部标识符 (Local Scope Identifier, 32 bit)，仅在局部网络范围内使用。

HIP 中并没有其他协议中定义的协议头部，而是用扩展头部来表示协议头部，用封装安全载荷 (Encapsulated Security Payload, ESP) 进行封装，在两个节点之间建立端到端 IPSec ESP 安全关联 (Security Association, SA) 来增强数据安全性，减少了中间节点 (如路由器) 对数据分组的处理，也不需要对现有的中间节点进行任何改动。

HIP 的报文一般由 HIP 头和 TLV 两个部分组成。一个 HIP 报文必须包含一个 HIP 头，可能包含一个或多个 TLV，也可能不包含 TLV。

HIP 定义的 HIP 头部结构如图 24 所示。

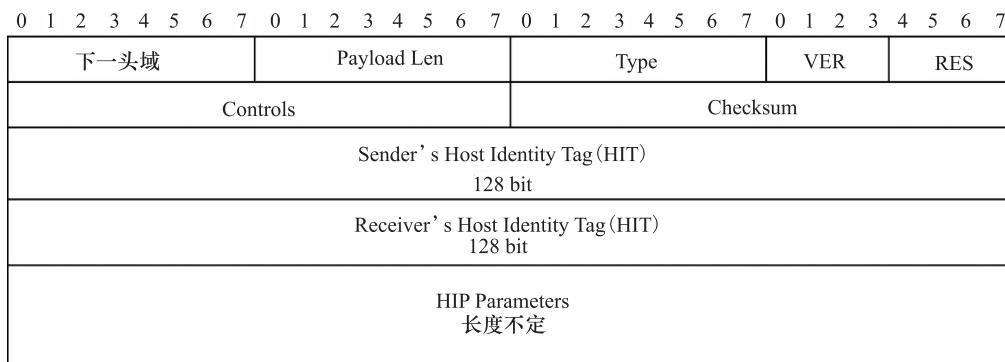


图 24 HIP 头部结构

其中，HIP 头部本身就是 IPv6 的一个拓展头，因为目前 HIP 扩展头为 IPv6 扩展头中的最后一个，它后面不应该再跟其他 IPv6 扩展头，所以 HIP 头部中下一头域 (Next Header) 的取值为 59，即目前的 IPv6 定义中表示为 IPPROTO_NONE 的数值。HIP 头部中的负载长度 = HIP 头部的长度 + HIP 头部后面的所有 TLV 的长度 - 1，单位为 8 Byte。Type 字段表示 HIP 报文的类型，如果收到的报文中的类型不能被识别，则必须丢弃该报文。VER 字段表示 HIP 的版本号，目前被暂时定义为 1。Controls 字段定义了一些 HIP 的控制信息。Checksum 域为校验和。HIP Parameter 字段填充的是 HIP 的 TLV 参数，即类型—长度—值这种类型的参数。RES 字段现在被保留以供将来扩充新功能之用，并且还包含了一个 128 bit 的发送方 HIT 和一个 128 bit 的接收方 HIT。

HIP 利用最简单的 DNS 实现节点标识到节点位置的映射。首先，一个域名被首先映

射为一组节点标识，节点标识再被映射为一组 IP 地址。

HIP RVS 机制可以为移动主机提供初始可达性。在 HIP 体系结构中引入了 RVS 服务器，节点移动后需要向 RVS 注册其节点身份标识符（HIT）和当前的 IP 地址。向 RVS 服务器注册后，RVS 的 DNS 域或其 HIPRVS，记录主机的 IP 地址。其他节点要与该移动节点通信时，首先查询 RVS 服务器，得知该移动节点的 IP 地址，然后将所有 HIP 报文发送给获得的 IP 地址。HIP 的切换过程可用下面的例子来表示：

- 节点 A 向 RVS 注册其 HI 和 IP 的映射关系。
- 当节点 B 要与节点 A 通信时，首先通过 DNS 查询得到其节点 A 当前的位置。
- DNS 返回节点 A 的 HI、RVS 及其 IP 地址信息。
- 节点 B 通过 RVS 向节点 A 发送 II 报文。
- 节点 A 与节点 B 建立基本连接的后续报文。
- 节点 A 移动到另一个子网中。
- 节点 A 向 RVS 注册其地址更新，更新 HI 和 IP 地址间的映射。
- 节点 A 与节点 B 继续通信。

HIP 的优点在于其在设计之初就在协议级别将节点身份和网络位置标识区分开，HI 或 HIT 用来标识节点身份，IP 地址仅用来标识网络位置。其优点体现在对移动性的支持上，因为使用了 HI 或者 HIT，即使移动节点在网络中的 IP 地址不断变化，HI 或者 HIT 与 IP 地址的映射关系也能不断变化，从而保证了节点既保持连接，又不断移动。

应用 HIP 后，通过 HIT 进行通信会在通信的两端建立安全连接，所以把 HIP 应用到移动方面时，很大程度上会提高网络的安全性。它提供了基于加密节点标识的端节点认证，节点可以通过节点密钥对验证身份；HIP 还保证了数据报文和控制报文的完整和可信，控制报文可以携带加密证书，用于端节点和中间实体的认证。

HIP 的缺点在于：需要更新大量节点，不支持流量工程和多播，增加了中间设备的复杂性。RVS 服务的方式加快了节点标识的更新速度和映射信息的传输，但 RVS 服务器需要维护完整映射数据库，而节点标识名字空间十分巨大，这就加大了 RVS 服务器的实现难度。

2.4.3 LIN6

在 IETF 的第 49 次会议上，TeraokaF 等日本学者提出了一种全新的移动 IP——LIN6^[15]。LIN6 是根据 LINA，即基于位置无关的网络结构的原理，在 IPv6 地址中划分出身份和位置标识的部分，它面向 IPv6 提出了一种移动性支持方案。

LINA 秉承身份标识与位置标识相分离的思想，引入了接口位置识别号和节点标识号这两个基本实体，实现身份标识与位置标识相分离。这两个实体的引入，使得网络层被分为网络标识子层和网络转发子层，网络转发子层履行传统 IP 层的功能，为数据分组提供路由。

网络标识子层要完成节点标识号与接口位置识别号之间的相互转换。这个转换过程需要专门的映射设备，把不变的节点标识与可变的节点位置联系起来，在 LINA 中使用映

射代理（MA）。同 HIP 类似，节点移动时要及时向自己的映射代理更新自己的映射。为了与传统协议兼容并且简化协议本身，LINA 采用了嵌入地址模型方法，即把节点标识号嵌入接口位置识别号里，得到新接口位置识别号，称为 ID—嵌入位置识别号。

LIN6 的基本思想是：采用 ID—嵌入位置识别号，将 IPv6 地址分为身份标识（LIN6 ID）和交换路由标识（LIN6 前缀）两部分。LIN6 ID 在上层应用标识通信，身份到路由的解析由终端的协议栈与映射代理通信来实现。与 HIP 一样，同样使用 DNS 将节点与其对应的映射代理服务器联系起来，通过部署映射代理服务器（MA）实现身份标识和交换路由标识之间的解析。

LIN6 最初的设计目标就是改善 IPv6 网络的移动性，所以与传统的移动 IP 不同，LIN6 争取在协议的层面对节点的移动性进行优化，解决了传统移动 IP 的三角路由、开销过大、单点故障等一系列缺点。但是与 HIP 相比，LIN6 的安全性设计存在不足，只适用于 IPv6 网络；同时 LIN6 也有许多缺陷，如微移动（Micro-Mobility）切换时的时延较长，且有分组丢失、节点标识空间狭小等缺点。

2.5 其他新型网络体系结构

2.5.1 XIA

XIA^[16]是由波士顿大学、卡内基梅隆大学、威斯康星大学麦迪逊分校共同开发的一个开源项目，作为 NSF 未来网络结构研究第 2 阶段的 4 个项目之一，主要研究网络的演进，意在解决不同网络应用模式之间通信的完整性与安全性问题。

随着互联网应用的日益多样化，协调这些应用在互联网中进行通信的问题逐渐引起了关注。XIA 致力于解决端到端之间的安全通信，建立一个统一的网络，为端口间的通信提供接口（API）。由于网络的复杂性，在网络中运行的程序与协议具有不同的行为和目标，XIA 希望通过定义具有良好支持性的接口，让这些网络活动的参与者能够更有效地运行，消除网络基础架构与端用户之间的通信障碍。在构建统一的网络基础架构的思想上，XIA 通过其内部的机制实现安全性。运行在这个架构之上的所有网络活动参与者具有安全标识，并应用于信用管理中，称为“内在安全机制”。XIA 扩大了目前基于主机通信的机制，将互动机制应用于对主体（包括主机、服务、内容等）的操作以及安全控制，对网络的控制从单一的分组转发扩大到网络中的互操作。在保证安全性的基础上，XIA 提供了足够的可扩展性。由于 XIA 希望通过单一的网络结构实现对于安全性的控制，必然需要提供演进的能力以支持不断出现的新的应用。以网络实体为例，从最初的主机发展到目前以内容为中心的趋势下出现的服务、内容主体以及未来可能出现的主体，XIA 提供灵活的绑定机制支持这些主体通过接口连入网络。

XIA 有 3 个关键的理念：①丰富的通信实体集合，XIA 的网络体系结构本质上支持不同实体间的通信，包括主机、服务、内容和其他未来使用模型中出现的实体；②内在的安全性，对所有实体使用标识符，并支持系统性的认证机制；③无处不在的“细腰”

“细腰”模型是指上层应用与底层链路之间具有的较小的协议中间层，其简洁性的优点使互联网快速发展起来，但是目前这一模型遇到了瓶颈。XIA 基于现有 Internet 的“细腰”模型，在安全性与扩展性方面进行了改进：第一，对于所有网络主体的支持性，XIA 为所有类型的主体定义了与不同协议机制之间的接口；第二，增强了信任管理；第三，保持“细腰”结构的简单性，同时将地址标识替换为服务标识。

XIA 的体系结构如图 25 所示，图中显示了 XIA 的组件以及相互之间的关系。XIA 的核心是最底层的协议（XIP），可以支持多种类型的通信实体间的通信。根据主体的操作目的对三种主体类型进行了不同的定义，内容被定义为它是什么，主机被定义为它是谁，服务被定义为它做什么。不同类型的主体需要定义各自的服务标识，例如，在基于内容的网络中，需要提供 API 来供用户获得、发现和搜索内容。XIA 使用“细腰”模型定义互操作需要的最小功能，该“细腰”不要求实施的精确过程，因此网络可以根据角色的类型来确定通信的类型。同时 XIA 的设计还支持未来的其他实体，例如用户和组。

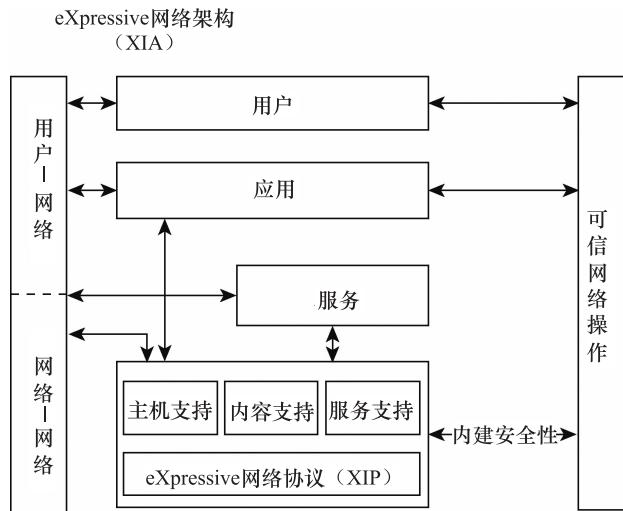


图 25 XIA 体系结构

网络主体之间的通信通过内容标识进行。它是一个 160 bit 的标识符（XID），可以表示一台主机（HID）、一条内容（CID）或者一项服务（SID）。这一标识具有安全验证的功能，利用公钥或者散列校验的方法，无需依赖外部的数据库就能进行验证，因此安全特性是内建的。当需要数据时，数据接收方能够获取想要的数据，并验证其来源。

以用户访问 Web 为例，用户将 URL 提交给 HID 服务器，这里的 HID 可以是 URL 的一部分，也可以是可信方式获得的主机名。服务器返回一个或多个内容 CID，用以标识页面的内容。用户的浏览器检索服务器、缓存或者网络中的内容副本，以获得相应 CID 的内容。

XIA 的子项目，如 Tapa、DOT、SCION 等，研究数据传输层面的可扩展性与安全性。Tapa 项目作为基础，主要解决异构网络上分组传输的体系结构问题。DOT 项目深入研究面向数据的传输和缓存问题。SCION 项目基于路由控制和故障隔离的目的，将独立的 AS

划分为可信域，并进行互联，以形成完整的传输路径，在受到攻击或者发生故障时，网络能够采用弹性的机制来保证故障的恢复。

除目前 XIA 已经提供基础的体系结构，还有一些正在开发的基于 XIA 的项目。支持复杂的通信主体类型以及所有通信操作相关的内在安全属性，是 XIA 两个特性。XIA 的特性也可以用于支持不同类型的移动性，例如，机器间的进程迁移或设备在网络中的移动性。一个关键的挑战是确保协议的安全性，例如，协议不能被第三方劫持，以便用于开放的通信会话。如何平衡用户隐私与网络管理的有效性和可管控性之间的矛盾，是网络体系结构需要解决的挑战之一。XIA 一直在探索使用“隐私按钮”的方式，用户将可以通过单一的按钮，获取诸如 ISP 等信息对通信进行控制，如避免某些类型的 XID、自动调用类似 TOR 的匿名服务等。由于接口在应用程序和协议栈之间，因此这一机制是跨应用的。

XIA 预期的未来互联网络模型是一个单一的网络，这与现今的互联网不同，它着眼于安全性的问题，支持网络的长期演进。原有的网络主要是基于主机的通信，而现在越来越多的应用的目的是内容获取，这也是 XIA 设计的一个挑战。未来的互联网不能只支持现在流行的通信主体（主机和内容），而必须是灵活的、可扩展的，才能支持互联网使用过程中出现的新实体。对于不同的网络主体，XIA 需要提供与主体特性相适应的属性和协议。互联网中的实体拥有不同的操作目标，XIA 支持网络角色的显式接口，网络主体需要的 XID 是由系统的协议给出的，XIA 设计不同的机制以适应不同的网络主体。另外，XIA 还对用户与网络、网络与网络的通信进行了区分，为两者设计了不同的接口。XIA 探索不同的标识符栈的组织方式和不同的分组路由，以便寻找不同机制来支持不同范围的网络服务，这涉及服务标识符的定义、粒度的控制、缓存以及内容分发等，功能涵盖端到端传输、分组转发、内容和服务支持，同时还要对这些操作进行可信管理。这些探索的目的在于支持长期的技术演进。随着链路技术以及存储计算能力突飞猛进的发展，网络体系结构必须支持新技术的高效整合，才能适应技术进步和经济发展。

2.5.2 NEBULA

NEBULA 是一个具有内建安全性的未来互联网体系结构，在满足灵活性、可扩展性和经济可行性的同时，可以解决新兴的云计算的安全威胁问题，其核心是一个高度可用、可扩展的由数据中心构成的网络。

2.5.2.1 NEBULA 的研究目标

NEBULA 项目为该体系结构设定了以下研究目标。

(1) 安全性与可信性

一个新的互联网需要超越可用性与健壮性，确保用户数据的安全和保密以及数据传输路径的可信与保密。NEBULA 数据平面可以解决这些问题，它能保证数据传输过程中的路径可靠，在传输过程中数据不被篡改，并且是保密的。在用户数据的迁移过程中，数据需要从一个数据中心迁移到另一个数据中心，跨越不遵循相同路由策略的网络，因此数据的超长距离通信也是 NEBULA 数据平面需要解决的问题之一。数据传输的路径上

需要有联合控制策略，数据中心的计算与存储需要隔离，更进一步地，操作系统和网络使用统一认证和授权机制。

(2) 高可靠性服务和非破坏性升级

下一代网络设备需要有以下高可靠性服务特征：能连续运行，没有预定的停机、日常维护和重启的时间，系统要能够承受彻底的攻击，有可靠的硬件，进行软件冗余备份，准备热备件，有完善的快速恢复计划。为了实现服务提供者改变服务或者在不删除旧服务的情况下部署新的服务，系统需要支持虚拟化。正如云服务供应商支持程序代码多个副本的同时运行，未来的网络设备供应商也需要支持多个路由协议副本无干扰地并行运行。

(3) 整合数据中心和路由器

现代的核心路由器是多个机架组成的大型分布式系统，NEBULA 研究了数据中心的计算机集群与核心路由器集成的可能性。为了保证可靠性，实现高吞吐量，在路由策略中允许并行转发路径存在。旧的路由协议无法满足对流量平衡的需求，无法实现数据中心与核心路由之间的最佳互联，因此对新的路由协议提出了要求。NEBULA 试图打破数据中心和互联网之间的屏障。

2.5.2.2 NEBULA 体系结构的组件

NEBULA 具有三个相关的组成部分：①NEBULA 数据平面（NDP），可以建立策略兼容的路径，提供灵活的访问控制并防御攻击；②NEBULA 虚拟可扩展网络技术（NVENT），它是 NEBULA 的控制平面，可以提供服务访问和网络抽象机制，例如冗余、一致、策略路由等；③NEBULA 核心，用于连接企业级数据中心的超可用下一代路由器。NVENT 使用策略可选的网络抽象来提供新的控制平面安全。NDP 使用创新的方法实现网络路径的建立，利用密码学机制在 NEBULA 路由器间建立策略可控的可靠路径。

(1) NEBULA 数据平面（NDP）

在 NDP 协议中，相对于分组路径中的每个管理域，分组具有下面 4 个元素：

- 域标识符。
- 管理域授权该路径的证明，称作 PoC。
- 分组遵循该路径的证明，称作 PoP。
- 类 MPLS 标记。

这个标记是连接已认可的通信与策略相关数据平面功能的函数，它可以映射 RBF 风格的规则，提供 RBF 项目的功能和灵活性。这个标记也可以表达查询优先级，限制域内路由，授权中间盒或者流量整形，或其他未来的数据平面特性。

这 4 个元素对于策略的表达和执行都是足够的。当分组到达管理域时，域拥有决定是否为该分组分配内部资源的全部信息，例如，分组是否被授权（检查 PoC），分组需要消耗哪些内部资源，要穿过哪个中间盒（检查标记），分组是否遵循授权的路径（检查 PoP 标记）。

初步的试验和原型系统表明，该体系结构在分组空间和数据平面处理成本两方面都是可行的。对比先前的工作，即使在强威胁的模型下，NDP 也可以真正强制执行策略目

标。特别地，NDP 提供以下性质。

- **路径保证：**对于即将发生的通信，沿路径的所有实体必须全部认可该路径。这个特性归纳了先前的许多成果，例如发送者控制下行路径，提供商控制前一条和下行路径等。
- **访问控制：**当路径不被认可时，分组不会被转发。由于路径包括目的地和潜在的服务标识符（称作目的地的标记），该体系结构可以简洁地实现访问控制，该功能通常被称作“将防火墙放入网络内”。
- **可用性：**由于路径选择发生在数据平面外，终端节点有充足的机会协商多条路径，如果路径失效，则启用备份路径。根据评估，这个过程远快于 BGP 计算新路由所需的时间。
- **自主资源控制：**任何实体都不会被强制以自己不赞同的方式部署自己的资源。该特性是安全的基本组件，可以确保不会违反任何实体的传输策略。
- **保护隐私的通信：**隐私包括通信内容的保密和该通信事件的保密。前者是网络层之上的问题，而后者是网络层应当考虑的，NDP 支持两个通信实体控制通信如何进行。通信实体可以使分组通过自己信任的提供商，更大胆的做法是端点指定一个“洋葱”路由系统，或者指定通信实体间的通信必须通过一个隔离的信道。

(2) NEBULA 虚拟和可扩展网络技术 (NVENT)

现有的互联网是以企业为中心的，这假设不同的组织分别运行服务器，通信发生在作为端点的独立的计算机上。与此相反，云是以服务和数据为中心的：计算和数据服务可以由多个数据中心冗余地提供，数据中心可以相互复制来提供更高的可靠性和性能。云允许多样的进化，例如面向内容的云。NEBULA 是一个具有进化能力的网络体系结构，只用在高可用性需求的服务所在地提供一个新的核心 (NCore) 就可以轻松扩展 NEBULA。当新的服务在路由器上可用时，NVENT 会发现它们。

(3) NVENT 对分布式服务的支持

NEBULA 研究的一个方向是对移动用户和分布式服务提供更好的网络支持，实现方法为：把人类可读的主机名转变为机器可读的服务标识符；将独立的分组转变为流；将单播通信转变为任播通信。NEBULA 实现移动性的方法是对应用隐藏网络地址，随着端点的改变（例如虚拟主机迁移、故障、设备移动等）进行动态重映射；紧密整合服务端点和网络元素，以提供更好的可扩展性和对变更的响应。

此外，一个服务实例可以托管于多个机器上（通常称为 Shard），这需要高度可靠的域内和域间的路由协议。这些协议必须能够反映真实世界的商业关系，并且只要存在一条策略兼容的路由就能够保证流量转发。当资源不能通过路由回传时，这条路由是无用的，需要进一步地改变互联网资源发现和资源分配的性质，才能保证即使在攻击者使用拒绝服务或路由劫持来阻塞访问时数据分组仍然可以顺利投递。NDP 规定了数据平面的机制，NVENT 规定了控制平面的策略框架，但还需要一个在域内和域间层级上的具备快速恢复能力的分布式状态管理框架。

NVENT 服务接口允许应用程序或接入提供商发送服务请求，并说明需要的可用性等

级。例如，一个提供应急服务的接入提供商可以要求具有多径域间路由的高可用性，实现该方案的关键是灵活性：每种服务可请求适当的传输，而不局限于某个单一版本的特性集合。NEBULA 设想使用分布式的解析服务来提供每个服务的信息、访问服务的方法以及服务的性质。这种全球规模的解析服务将提供超越 DNS 和 BGH 系统的可扩展性、灵活性与动态性。

(4) NVENT 和 NDP 的接口

NVENT 的工作是决定分组元素的合适值。NVENT 负责决定分组路径，收集所有中间域的许可（PoC）并确定路径中的标记。通常情况下，预期的发送者向 NVENT 服务器查询，并将这些信息放入分组内。当 NDP 分组进入网络内时，沿途的域有足够的信息进行检查。

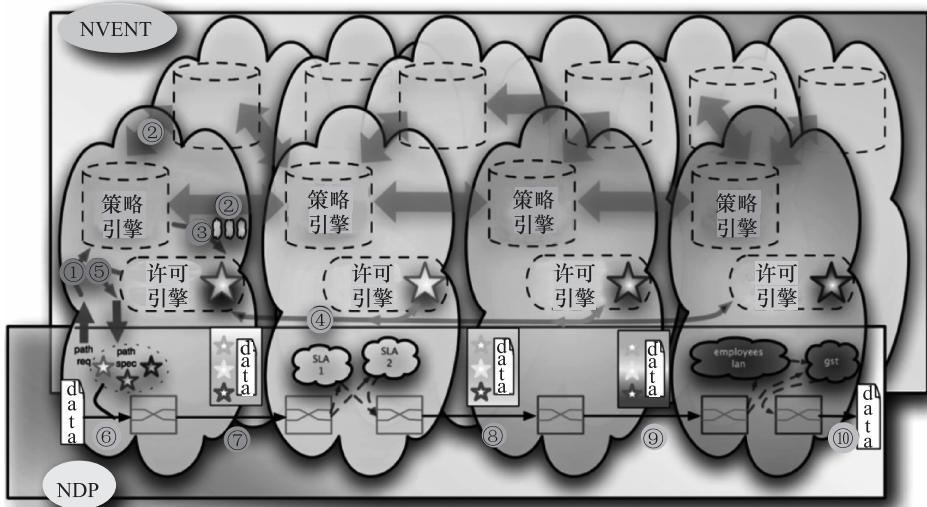


图 26 交互过程

(5) 责任机制

责任机制是减少故障和不当行为的有效方法。故障的原因是多样的，可能是意外的配置错误，甚至是有意的攻击。责任机制保证故障和错误行为被有效地监测，并向网络管理员提供故障的快速反馈，即便系统不能阻止或消除它们的影响。责任机制也能定位故障所在的组件或域，使得每个域可以相互追究责任，进而激励每个域保持自己的设施尽量可靠。

(6) NVENT 控制策略

为了简化用户控制，NVENT 使用声明式网络作为 NVENT 的网络配置框架。声明式网络是一种有助于开发者精确地说明网络协议和功能的编程方法，网络协议和服务可以被编译为严格执行说明规范的硬件指令。NVENT 计划开发 NDLog（Network DataLog）语言，该语言的特点是：允许用户高效地描述并构建灵活的网络服务和 NDP 分组规则；具有一个高效的编译器将 NDLog 语言转换为底层网络指令（例如 OpenFlow 交换机的配置）。

2.5.2.3 NEBULA 核心

NEBULA 核心将会构建在未来核心路由器的基础上，它能够支持在任意时间内保持最高的传输速率，同时保持 always-on 级别的可用性。

单一 CPU 不能满足 Tier-1、ISP 转发速率的需求，这要求下一代路由器的控制平面设计为一个容错的分布式系统。下一代路由器包括多个机架（思科公司计划在近期将路由器规模扩展到多达 48 个机架），每个机架有多个线卡、转发处理器和控制处理器。现有的互联网协议无法很好地处理网络中的交互操作，实现语义精确、快速故障切换、连续运行。越来越多的网络服务要求高可用性和对事件的一致性响应，包括路由更新、管理命令和服务请求。为了支持分布式安全架构和可信的核心路由，路由体系结构需要重新设计，以增加其可扩展性，适应高可用性的要求。

NEBULA 对数据中心和核心路由进行了整合，创建了一个包括硬件和软件在内的体系结构，将数据中心和核心路由器直接相连。NEBULA 小组和思科公司以及英特尔公司合作，建立了一个计算集群和核心路由高度连接的网络。希望通过这一研究能同时实现高速率和高可靠性，以解决目前存在的数据存储和计算之间速率不匹配的问题。

为了可靠地运行分布式路由系统，NEBULA 设计了新的软件栈和新的动态可重构服务（DRS）模型，为路由器提供一致性模型。路由器需要分布式的，并且能够在不停止服务的情况下升级，一致性模型能够对路由器的操作属性进行复制，构建可靠和安全的分布式路由系统。

路由系统建立以后，需要对其可靠性进行监测。监测的引入会降低控制平面路由器的性能，但另一方面，也是快速故障恢复中不可或缺的一部分。NEBULA 设计了一个故障监测的自检系统，包括硬件和软件的全局视图、能够检查数据分组以及转发表，并且在数据面和控制面都有一定的处理权限，如可以创建一个时间标记数据分组、检查转发路径和时延等。

2.6 典型新型网络试验床

2.6.1 PlanetLab

2002 年 3 月，Larry Peterson（普林斯顿大学）和 David Culler（加州大学伯克利分校和 Intel 研究院）组织了一个在全球范围内对网络服务有兴趣的研究人员会议，提议将 PlanetLab^[17]作为研究团体的试验平台。这个由伯克利—Intel 研究院主持的会议吸引了 30 名来自 MIT、华盛顿、莱斯、普林斯顿等大学的研究人员。在随后的几年，该项目得到学术界、产业界和政府机构的广泛参与。PlanetLab 是用作计算机组网与分布式系统研究试验床的计算机群。它于 2002 年设立，到 2006 年 10 月由分布在全世界 338 个站点的 708 个节点组成。它是一个开放的、针对下一代互联网及其“雏形”应用和服务进行开发和测试的全球性平台，是一种计算服务“覆盖网络”（Overlay），也是开发全新互联网

技术的开放式全球性测试平台。每个研究项目有一个虚拟机接入节点构成的子网。在这之后的几年时间里，学术界、产业界和政府广泛地参与了此项目。截至 2009 年 6 月 3 日，PlanetLab 拥有 1 006 个节点和 475 个站点。它是一个开放性的、用于研究下一代互联网的全球性开发测试平台。

PlanetLab 最初的核心体系结构由普林斯顿大学的 PetersonL、华盛顿大学的 AndersonT、英特尔的 RoscoeT 以及负责此项工作的 CullerD 共同设计。PlanetLab 本质上是一个节点资源虚拟的覆盖网络，一个覆盖网的基本组成包括：运行在每个节点上以提供抽象接口的虚拟机；控制覆盖网的管理服务。为了支持不同网络应用的研究，PlanetLab 从节点虚拟化的角度提出了“切片”概念，将网络节点的资源进行了虚拟分片，虚拟分片之间通过虚拟机技术共享节点的硬件资源，底层的隔离机制使得虚拟分片之间是完全隔离的，不同节点上的虚拟分片组成一个“切片”，从而构成一个覆盖网。各个切片之间的试验互不影响，而使用者在一个切片上部署自己的服务。

PlanetLab 的架构如图 27 所示，每个节点通过 Linux vServer 虚拟机技术虚拟成多个 Silver，不同节点的 Silver 形成一个 Slice（即虚拟网络）。使用者在一个 Slice 上部署自己的服务，各个 Slice 之间的试验互不影响。研究人员能够请求一个 Slice 用于试验各种全球规模的服务。目前在 PlanetLab 运行著名的服务主要有：CoDeeN 和 Coral CDN，ScriptRoute 网络测量服务，Chord 和 OpenDHT，PIER、Trumpet 和 CoMon 网络监控服务。

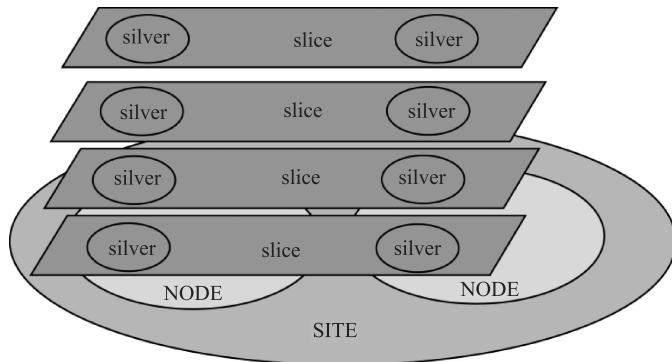


图 27 PlanetLab 系统框架

PlanetLab 的主要目标之一是用作重叠网络的一个测试床。任何考虑使用 PlanetLab 的研究组都能够请求一个 PlanetLab 分片，在该分片上试验各种全球规模的服务，包括文件共享和网络内置存储、内容分发网络、路由和多播重叠网、QoS 重叠网、可规模扩展的对象定位、可规模扩展的事件传播、异常检测机制和网络测量工具。

PlanetLab 的优点在于它的节点是真实地分布在全球的各个地方，研究人员可以部署真正意义上全球范围的试验应用。而且，PlanetLab 上运行的试验有效的运行周期是 2 个月，用户可以观察试验长期的运行结果，以有效地评估试验的前景。PlanetLab 上提供了

一套名为 MyPLC 的软件，用户通过在自己的节点上安装这个软件加入 PlanetLab 中，成为 PlanetLab 的一个站点。PlanetLab 的不足之处在于，普通的试验用户对于 PlanetLab 上的资源只有部分的 root 权限，他们只能在节点上部署应用层的试验，无法进行底层的网络技术研究。PlanetLab 系统框架如图 27 所示。

PlanetLab 也可以作为一个超级测试床，在其上有更多的狭窄定义的虚拟测试床能够被部署，即如果将服务的概念泛化（一般化）以包括传统上认为的测试床，那么多个虚拟测试床能够在 PlanetLab 上部署。例如，正在开发一个“分片中的 Internet”服务，其中在一个分片中重新创建 Internet 的数据平面（IP 转发引擎）和控制平面（如 BGP 和 OSPF 的路由协议）。网络研究人员能够使用这项基础设施来进行 Internet 协议簇的修改和扩展的试验。

除了支持短期试验外，PlanetLab 也可以用来支持长期运行的服务，这些服务支持一个用户基础（用户群）。与其将 PlanetLab 严格地看做一个测试床，不如采取更长远的观点，将其看做既是一个测试床又是一个部署平台。因此，PlanetLab 支持一个应用的无缝迁移，从早期原型，通过多次设计迭代，到一项持续演进的受欢迎服务。

由于 PlanetLab 节点遍布世界各地，因此一个 Slice 上的虚拟机也就遍布世界各地，这样用户得到了一个由遍布世界各地的服务器组成的网络。在这个网络上，用户可以进行全球范围的、真实环境下的网络试验。

2004 年 12 月 27 日，中国教育和科研计算机网（CERNET）加入 PlanetLab，CERNET 的加入标志着 PlanetLab 中国项目的启动，CERNET 首先在中国 20 个城市的 25 所大学中设立了 50 个 PlanetLab 节点，这使得 CERNET 成为亚洲第一个地区性 PlanetLab 研究中心。

PlanetLab 由一个管理中心（PLC）和遍布全球的几百个节点组成。一个节点就是一台运行着 PlanetLab 组件的计算机（服务器）。节点由许多独立的站点管理和维护。这些站点包括大学、研究机构和 Internet 商业公司等。一般来说，每个站点至少提供 2 个节点的服务。每个节点上同时运行大量的 Sliver，节点的资源（包括 CPU 时间、内存、外存、网络带宽）被分配给这些虚拟机。虚拟机如同 Internet 上的真实主机一样，可以安装和运行程序。

由许多节点上的虚拟机条带组成的一个环境叫做 Slice。用户在 PlanetLab 上的试验部署在各自拥有的切片上，也就是部署在由每个节点上的一个虚拟机组成的一个大规模网络试验环境上。由于 PlanetLab 节点部署在世界各地，因此切片网络上的虚拟主机也就遍布世界各地，这样用户就获得了一个由遍布世界各地的主机组成的网络试验环境。借此，用户可以进行全球范围的、真实环境下的网络试验。PlanetLab 的这套设计思想被研究者称为基于切片的计算（Slice-Based Computing）。

所有 PlanetLab 机器都运行一个常规软件包，包括一个基于 Linux 的操作系统、启动节点、分发软件更新的机制、监控节点健康、审计系统活动并控制系统参数的管理工具集、管理用户账户和分发密钥的工具。PlanetLab 的体系结构如图 28 所示。

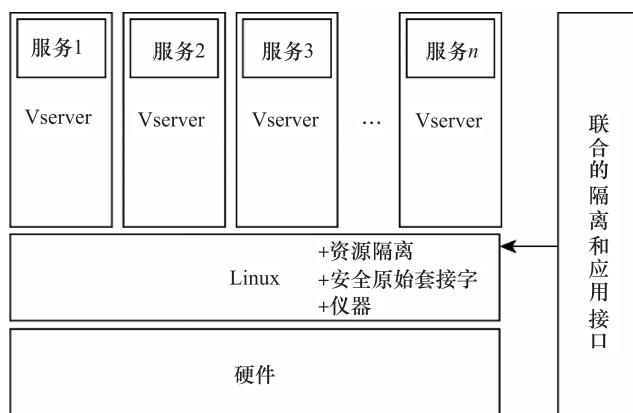


图 28 PlanetLab 的三层体系结构

2.6.2 GENI

全球网络创新环境 (The Global Environment for Network Innovations, GENI)^[1] 是美国 NSF 资助的一个关于下一代互联网研究中的重大项目，该项目旨在为未来的网络技术研究提供一个统一的网络试验平台。GENI 由一系列网络基础设施组成，可以为研究者提供大规模的网络试验环境，能支持多种异构的网络体系结构 (包括非 IP 的网络体系结构) 和深度可编程的网络设施。

GENI 的目标是构建全新的、安全的、灵活自适应、可与多种设备相连接的互联网，搭建基于“SourceSlice”有效调度的试验网络，为不同的新网络方案搭建试验平台。大部分新型网络体系都可以部署在这个试验平台中，从而达成一个物理网络支撑多个逻辑网络的目标。

GENI 以时间片和空间片的形式提供网络资源的虚拟化，使用户有机会创建自定义的虚拟网络并开展相关试验，从而摆脱现有互联网的一些限制。一方面，假如资源是以时间片的形式进行分割的话，可能会出现用户的需求量超过给定的资源，影响了其有关可行性的研究；另一方面，假如资源是以空间片的形式进行分割的话，则只有有限数量的研究者能够在他们的切片中包含给定的资源。因此，GENI 提出了基于资源类型的两种形式的虚拟化来保持平衡性，也就是说 GENI 采用时间切片的前提是有足够的容量支持部署研究。GENI 提供的虚拟网络能承载终端用户的真实网络流量，并连接到现有的互联网上以访问外部站点。

GENI 项目的目标是创建一个新的互联网和分布式系统，其具体目标包括以下 5 个方面：

- 具备安全性和顽健性。GENI 专家认为，重新考虑互联网设计的一个重要原因和动力是极大地提高网络的安全性和顽健性。目前 Internet 在网络安全方面的支持较差，尽管存在许多安全机制，但缺乏一个完整的安全体系结构，无法将这些安全机制组合起来为用户提供全面良好的安全性能。

- 实现普适计算，通过手机、无线技术和传感器网络更好地连接虚拟和真实世界。
- 控制并管理其他重要网络基础设施。
- 具备可操作性和易用性。
- 支持新型服务及应用。

GENI 项目是一个规模庞大、结构复杂、需求多变的工程。项目实施和完成，必须有一个好的设计原则做支撑。好的设计原则为满足未来互联网在安全、QoS 等方面需求，提高 GENI 设计寿命提供重要保障。为保证 GENI 对控制性试验和长期配置研究的顺利进行以及满足大面积分布式计算的需求，GENI 必须满足如下条件：

- 项目设计要有优秀的系统体系结构，项目建设要有选择性。
- 项目所有设计必须是开放的。
- 能为保证不同研究方向的研究团体能够共享 GENI 资源，需通过虚拟化或分割（时分或空分）技术将 GENI 资源划分为不同功能、相对独立的资源子集，从而保证研究团体能够顺利开展工作。
- 通用性是系统能够广泛应用的基础，同时系统也要有较强的安全性和顽健性。
- GENI 要有可访问性，能为用户提供同 GENI 连通的物理连接，也能为用户的加入提供多种连接机制。允许试验持续进行，也支持 GENI 同传统网络的连接。
- 为满足当今和将来用户的需求，GENI 需在无线技术、光技术、计算技术等方面取得发展和突破，并利用这些新技术探索新的应用和系统，给用户带来更方便、快捷的服务。
- GENI 提供的功能必须同现实中的事务或功能相吻合。
- GENI 具有多样性和扩展性，能对未知网络和网络新技术提供足够支持；同时也要有继承性。要继承现有网络技术中的优秀成果，利用现有网络基础设施，以现有软件及相关技术为平台进行 GENI 研究。这样，在降低项目投入的同时，也实现了同传统网络的平滑过渡。
- GENI 要有强大的隔离性，从而保证在某些切片出现故障时，不对其他切片产生影响。在网络管理方面，要求所有网络平台具有报错功能，能利用顶层协议描述和配置所有网络区域。当网络出现故障时，能提供诊断、反馈问题和报告错误的工具。
- GENI 在广泛部署的前提下，能够利用一定手段对 GENI 的相关性能参数进行测量，并对其进行量化研究。
- 从用户角度来说，GENI 在提供易用性的同时，也要保障资源不被攻击或窃取。

GENI 的主要设计原则包括以下两方面。

- 可切片化：为了提高效率，GENI 必须能同时支持多个不同使用者的试验，而虚拟化是达成这一目标的关键技术，其将在时间和空间上对资源进行划分。
- 通用性：GENI 为研究者提供了灵活的试验平台，这就要求平台的组件是可编程的；其他要求还包括支持广泛的接入技术和互联、接口标准化（可扩展性）、多级别虚拟化（组件重用）以及切片之间的隔离等。

GENI 的整体架构如图 29 所示。下面简单地介绍几个关键概念。

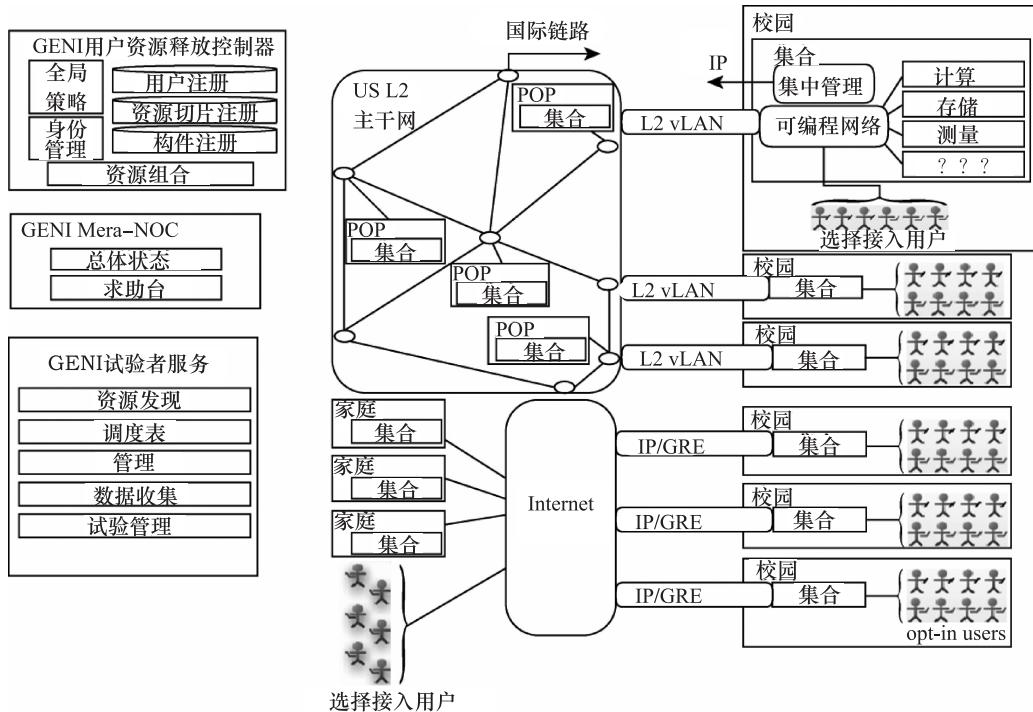


图 29 GENI 框架

- Component: 网络中的物理设备，例如路由器、交换机、物理链路等。
- Aggregate: 一个区域内 Component 的集合，在 GENI 中典型的 Aggregate 就是各个高校中负责的试验网络。
- Slice: GENI 中的 Component 通过虚拟化技术进行资源切片，资源片组成的虚拟网就是一个 Slice。
- Clearinghouse: GENI 的管理系统，负责管理用户注册、网络设备注册、虚拟子网注册等。
- Meta-NOC: GENI 的网络测量系统，负责测量和监控整个网络的状态。
- Experiment Services: GENI 为试验用户提供支持服务，比如资源的发现和调度、试验数据的采集、试验项目的管理，这为研究人员试验 GENI 提供了方便。
- Opt-In Users: 选择接入 GENI 的终端用户，他们是一些受 GENI 信任的终端用户，负责体验研究人员部署在 GENI 上的试验。

GENI 的发展思路是先由一些高校各自负责一部分网络实验平台的建设，称为 GENI 的一个簇。目前 GENI 由 4 个簇组成，它们分别是普林斯顿大学负责的 PlanetLab、犹他大学负责的 ProtoGENI-Emulab、杜克大学负责的 ORCA-BEN、罗格斯大学负责的 ORBIT-WINLAB。GENI 的这些簇通过 2 层的 VLAN 技术或 GRE 等隧道技术与 Internet 2 连接起来，组成整个 GENI 底层网络（Internet 2 是美国用于下一代互联网技术研究的一个试验骨干网）。

GENI 采用软件工程中的螺旋模型进行开发，这种模型的每一个周期都包括需求定义、风险分析、工程实现和评审 4 个阶段。整个开发工程由这 4 个阶段循环迭代完成。螺旋模型的优势在于它是一个不断迭代的过程，在每个为期不长的迭代周期中发现设计和实现中的漏洞和风险，并予以改进。目前 GENI 处于第 3 个螺旋，它包括大量的子项目，取得了一系列重要的成果。其中 PlanetLab 和 ProtoGENI 专注于 IP 网的研究，而 ORCA 和 ORBIT 则关注无线网的技术研究。

GENI 为网络虚拟化的研究提供了一些有意义的指导思想，GENI 认为网络虚拟化环境下的网络设施应该有以下特点：

- 可编程：研究人员可以在网络中的节点上部署自己的软件，控制这些节点的行为。
- 资源共享：网络设施可以同时并发地支持多个试验，不同的试验是隔离的，不会相互影响。
- 切片式管理：切片是试验所用的虚拟机节点和虚拟链路的集合。实验平台的管理系统以切片为单位管理整个网络中的物流资源。
- 联盟化：GENI 中的组件可以由不同的组织负责，这些组织共同构成 GENI 的生态系统。

GENI 设施的本质是能够快速、有效地嵌入一个大规模试验网络中，与其他设施和现有互联网相连提供网络运行环境，并且研究者可以通过严格观察、测量，记录试验结果。实现这些功能需要 GENI 设施跨越各种现有和未来的技术、网络架构、地理延伸和应用领域。

2. 6. 2. 1 系统架构

GENI 的体系结构可分为 3 层，自上而下分别是：用户服务层、用户管理核心 (GMC) 层和物理层（如图 30 所示）。GMC 层通过设计可靠、可预测、安全的体系结构，利用抽象、接口、命名空间同 GENI 体系结构绑定起来。考虑到物理层和用户服务层具有动态变化性，为快速、高效同其连接，GMC 定义了一套瘦腰机制，使其既能支持和适应物理层和用户服务层的发展，同时也能独立发展，以适应 GENI 整体发展需求。物理层通过提供物理链路，使用物理设备（如路由器、处理器、链路、无线设备等）实现网络内部节点之间的互联互通。用户服务层则通过提供服务访问接口，实现用户对 GENI 的访问。同时用户服务层具有可扩展性，能让服务在其生命周期内不断发展。

类似当前的 Internet 体系结构，GENI 体系结构类似沙漏模型，如图 30 所示。GMC 对应 IP 层以及它的编址路由和服务模式，同 GENI 沙漏的腰部对应。高层的用户服务层

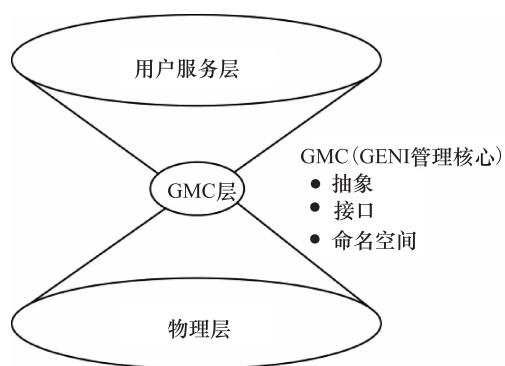


图 30 GENI 的三层体系结构

同那些附加的用于将 Internet 系统完整化的功能（如 WWW、Skype 等）相对应。GENI 底层对应着组成物理网络的计算设备和网络设备的集合。

物理层通过一定技术将一系列可扩展的组件组合起来，以满足用户社区的需求。图 31 描述了不同组件连接而成的物理层。从图中的看出，物理层由可编程边界链、可编程核心节点、可编程边界节点、客户端、全局光纤、微电路、多重网络交换节点、基于 IEEE 802.11 的城市无线子网、基于 3G/Wi-MAX 的无线子网和自适应无线子网构成。尽管这些组件不能单独运行，但 GENI 将这些组件组合起来构成虚拟网络，为研究者提供所需的试验条件。

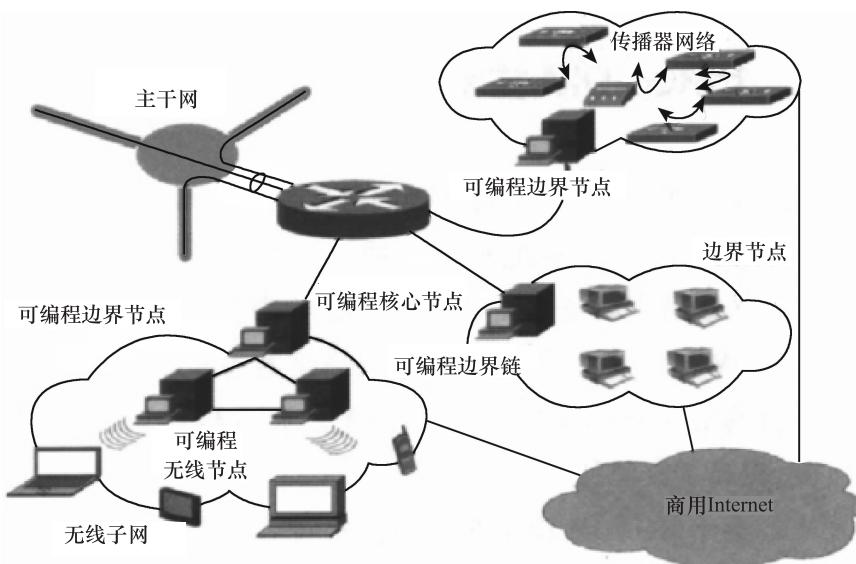


图 31 物理层结构

GMC 层通过一系列抽象、接口和命名空间同物理层相连，为上层用户提供服务。GMC 屏蔽了底层实现细节，为用户服务层提供相关信息。抽象是 GMC 层的关键，为屏蔽物理层细节提供了有效手段。GMC 层抽象分为组件、切片和聚合三种。组件是 GENI 的主要模块，包括物理资源、逻辑资源和同步资源，GMC 通过组件管理器，采用一定的组件协议将资源分配给用户；切片即相关 GENI 组件的切片，GENI 通过运行切片来实现用户需求，切片有效地保护 GENI 资源共享，保证了研究团体工作的顺利开展，有效降低了开发和运营成本；聚合是为实现某些组件和切片不能实现的特殊关系而提出的，是 GENI 中一个有效的补充。

用户服务层集中式地将模块组织起来同物理设备合并，从而形成一个能够支持研究的单一分布设施，以满足不同用户群体的需求。在物理层提供具体物理链路和 GMC 协调的情况下，用户服务层主要完成的功能如下。

- 允许拥有者为所控制的底层设备申请资源分配和使用策略，并提供确保这些策略实施的保障机制。
- 允许管理员对 GENI 底层进行管理。

- 允许研究人员创造和装配试验、分配资源并运行试验专用软件。
- 能将关于 GENI 底层的信息开放给开发者。

总之，GENI 的三层体系结构是一个有机整体，缺一不可。只有三层有机组合和相互协作，才能实现 GENI 的完整功能。

2.6.3 FIRE

2007 年，欧盟在其第 7 框架（FP7）中设立了未来互联网研究和试验（FIRE）^[49]项目。FIRE 的主要研究内容包括：网络体系结构和协议的新设计；未来互联网日益增长的规模、复杂性、移动性、安全性和通透性的解决方案；在物理和虚拟网络上的大规模测试环境中验证上述属性。对 FIRE 项目的发展，欧盟做了一个长期规划，初步将 FIRE 项目分为 3 个不同的阶段。目前 FIRE 项目进行到第二个阶段。在第一个阶段，FIRE 项目组一共支持 12 个项目，其中有 8 个项目用于试验驱动性研究，另外 4 个项目用于试验基础设施的建设；在第二个阶段，FIRE 项目组扩展了 FIRE 中试验驱动性研究和基础设施建设的项目，同时增加了一些协调与支持项目。通过对这些项目的研究，希望能够建立一个新的不断创新融合多学科的网络体系结构。FIRE 项目组认为未来的互联网应该是一个智慧互联的网络，包括智慧能源、智慧生活、智慧交通、智慧医疗等多个方面，这样就把社会中的各个方面通过互联网联系起来，最终实现智慧地球。

FIRE 和 GENI 有很多相似之处，它们都关注如何搭建试验环境为理论研究提供证据支持。GENI 也希望通过螺旋式的部署方案，突破地理限制，建立全球性的大规模试验环境；FIRE 同样采用虚拟化思想，将独立存在的资源和设施联系起来，也具有联盟和跨学科等特点。

FIRE 作为欧盟 FP7 在 ICT 领域的重要组成部分，是为应对未来互联网面临的诸多挑战而实施的大型研究计划，目标是逐步联合现有的和未来新的互联网试验床，建设一个动态的、可持续的、大规模的欧洲试验床基础设施平台，为欧盟互联网技术发展提供一个综合的研究试验环境。在 FIRE 中涉及试验床的项目有 4 个：OneLab2、PII、VITAL ++ 和 WISEBED。

OneLab2 基于欧洲 PlanetLab 试验床（PLE）平台，由 OneLab 试验床发展而来，在其基础上继续负责 PLE 的运作，并在网络监测、无线、内容网络、规范化测试等领域进行深入研究。

PII 建立在 PanLab 的基础之上，旨在开发高校的技术和机制来实现欧洲现有试验床的联合，从而建成一个超级试验床联盟平台。PII 联合试验床包括 4 个核心计算机群和 3 个卫星通信计算机群。

VITAL ++ 的主要目标是研究结合 P2P 和 IMS 各自优点的网络模型，并在试验床中进行试验和验证。按照 VITAL ++ 的设想，将用 IMS 技术把欧洲范围内的分布式试验床节点集合起来，组成一个 VITAL ++ 试验床。在这个试验床中，利用 P2P 技术进行内容应用和

服务试验，利用网络资源优化算法实现符合要求的 QoS；而整个试验床网络的管理、运行则通过传统电信网的方式实现。

WISEBED 计划将欧洲已有的试验床联合起来，目标是建设一个覆盖欧洲的、具有一定规模的无线传感网络试验床，为欧洲的研究者和产业界提供试验和服务。

2.6.4 AKARI 试验床

2006 年，在日本政府的支持下，新一代网络体系结构设计项目 AKARI^[2]在日本展开。AKARI 项目研究的是下一代网络体系结构和核心技术，分三个阶段（JGN2、JGN2+、JGN3）建设试验床，并在初期基于日本 PlanetLab 的 CoreLab。AKARI 研究规划从 2006 年开始，计划 2015 年完成，2015 年后通过试验床开始进行试验。

AKARI 是日本关于未来网络的一个研究性项目，AKARI 在日语中的意思是“黑暗中的一盏明灯”，它旨在建立一个全新的网络体系结构，希望能为未来互联网的研究指明方向。AKARI 的设计进程分为两个 5 年计划，第一个 5 年计划（2006~2010 年）完成整个计划的设计蓝图；第二个 5 年计划（2011~2015 年）在这个计划基础上完成试验台。在每个 5 年计划中，又对 AKARI 项目的进度进行了细分，整个项目的进度分为概念设计、详细设计、演进与验证、测试床的创建、试验演示等多个环节。AKARI 不仅对未来互联网整体架构进行设计，而且试图指明未来互联网技术的发展方向，希望通过工业界和学术界的合作，使新技术能够快速应用到工业化的产品中。AKARI 项目在设计时考虑到了社会生活中的各个方面，希望将社会生活中的问题和网络体系结构新技术的发展对应起来，形成一个社会生活和网络体系结构相对应的模型，希望网络新技术的发展与社会生活的需求相适应的。

在 AKARI 看来，未来网络的发展存在两个思路，即 NxGN（Next Generation Network）和 NwGN（New Generation Network）。前者是对现有网络体系的改良，无法满足未来的需要；后者是全新设计的网络体系结构，代表未来的方向。作为日本 NwGN 的代表性项目，AKARI 的核心思路是：摒弃现有网络体系结构的限制，从整体出发，研究一种全新的网络体系结构，解决现今网络的所有问题，以满足未来网络需求，然后再考虑与现有网络的过渡问题。AKARI 强调，这个新的网络体系结构是为人类的下一代创造一个理想的网络，而不是仅设计一个基于下一代技术的网络。

为此，AKARI 确定了设计需遵循的 3 个原则，介绍如下。

（1）KISS（Keep It Simple Stupid）原则

KISS 原则是指新的网络体系结构要足够简单，在具体研究中要贯彻以下 3 个基本理念：

- 透明综合原则：在选择和整合现有技术时要以简单为首要条件，剔除其过于复杂的功能。
- 通用分层的思想：新型网络体系结构要采用层次结构，各层功能要简单并保持独立性。
- 端到端原则。

(2) 真实连接原则

新型网络体系结构中，实体的物理地址和逻辑地址各自独立进行寻址，要支持通信双方的双向认证和溯源，确保连接的真实性和有效性。

(3) 可持续性和进化能力

新型网络体系结构应该成为社会基础设施的一部分，必须考虑今后50~100年甚至更长时间的发展需要，因而体系结构本身应该是可持续发展的、具有进化能力的。

图32是AKARI设计新型网络体系结构的时间表。AKARI计划2010年前完成体系结构的设计。在过去的4年中，AKARI研究了大量现有的技术方案，其中的15个方向是其新网络体系结构的重点研究内容，见表2。

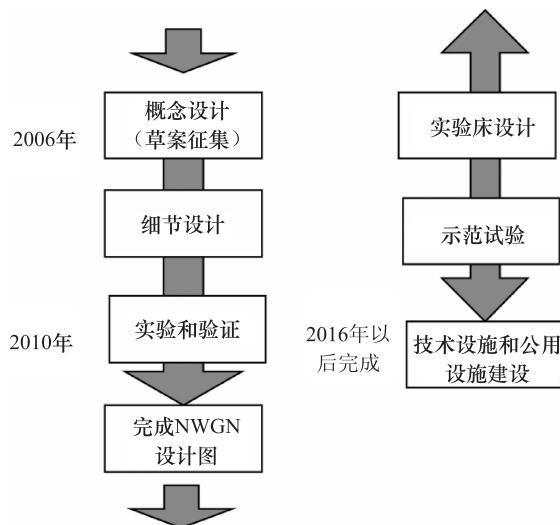


图32 AKARI设计新型网络体系结构的时间表

表2 AKARI计划的近期研究内容

光分组交换和光路技术	光接入	无线接入
分组分多址 (PDMA)	传输层控制	主机/位置标识网内分离体系结构
分层	安全	QoS 路由
新型网络模型	顽健控制机制	网络层次简化
IP 简化	重叠网	网络虚拟技术

目前，AKARI正在对上述内容进行研究，已经取得了不少进展，主要包括以下几个方面。

- 提出了主机/位置标识网内分离体系结构，和欧盟FR74WARD的WP6的设计思想类似，但走得更远，在2008年已经提出一种方案。
- 针对现有IP层越来越复杂的现实，AKARI提出了IP网络协议，以简化IP。
- 在新型网络体系结构引入最新的光网络技术，包括面向连接的光路技术和无连接的光交换技术，并且正在研究简化甚至去掉数据链路层的技术。
- 提出了穿越网络层次的控制机制，研究层与层之间交换控制信令，实现顽健控制。
- 与传统的7层网络模型不同，AKARI提出了基于用户的新型网络模型。

3 国内研究进展

我国也非常重视对未来信息网络体系结构和关键理论及技术的研究。“十一五”期

间，国家对新一代信息网络基础理论研究进行了重点支持。2006 年，国家 973 计划启动了“一体化可信网络与普适服务体系基础研究”项目^[22~24]，进行未来信息网络体系结构的基础研究。2007 年，国家 973 计划资助了“可测可控可管的 IP 网的基础研究”项目，主要针对现有 IP 网的可测可控可管性开展研究。2008 年，国家 973 计划资助了“新一代互联网体系结构和协议基础研究”项目，研究新一代互联网体系结构与协议。国家 863 计划信息技术领域 2008 年度专题课题资助了目标导向类课题“身份与位置分离的新型路由关键技术与实验系统”^[25]，主要研究身份与位置标识分离的新型路由寻址体系结构及解决方案。2010 年 11 月，国家 863 计划信息技术领域启动了“三网融合演进技术与系统研究”重大项目^[26]，将“面向三网融合的创新网络体系结构”列为重要研究内容。2011 年，国家 973 计划支持了“面向服务的未来互联网体系结构与机制研究”和“可重构信息通信基础网络体系研究”两个项目。前者以面向服务为核心设计理念，以服务标识作为沙漏模型的细腰，并以服务标识驱动路由和数据传输，在体系结构和核心机理层面进行针对性的研究；后者侧重于构建一个功能可动态重构的基础物理网络，为不同业务构建满足其需求的逻辑承载网，以解决目前网络层的功能瓶颈。2012 年 2 月发布的 973 项目申请指南将“智能协同网络理论研究”列为重要支持方向^[27]，并启动了“智慧协同网络理论基础研究”等项目。

3.1 开放可编程网络

围绕开放可编程网络，我国研究界和产业界做了大量的研究工作，并取得了一些成果。

清华大学于 2010 年同美国斯坦福大学签署了未来互联网和 SDN 的合作研究协议。首先从基于 SDN 的 IPv6 源地址验证入手。真实源地址验证是清华大学一项有特色的研究工作，通过保证地址的真实性来为网络的安全可信提供基础。基于 SDN 的源地址验证研究工作进一步拓展了真实源地址验证工作的应用范围，同时也展示了 SDN 的灵敏性对网络安全带来的进一步收益。相关成果除了在国际会议和期刊发表，也被收录在美国 2014 年 2 月出版的《Network Innovation through OpenFlow and SDN: Principles and Design》中的第三章《IP Source Address Validation Solution with OpenFlow Extension and OpenRouter》。

2013 年 1 月，国家“863”项目“未来网络体系结构和创新环境”启动。该项目由清华大学牵头负责，清华大学、中科院计算所、北邮、东南大学、北京大学等分别负责各课题，项目首席专家为清华大学毕军。项目的出发点在于：新型网络体系结构和新型网络协议在设计方法、编址方式、转发机制、控制模式等方面存在巨大差异，而当前网络设备和试验环境的封闭性又严重制约着网络的技术创新和体系演进，因此，设计和实现促进未来网络体系结构创新的环境，对支撑未来网络的技术创新和体系演进具有重要的意义。该项目的目标是：采用创新理念和技术路线，研究未来网络创新体系结构，突破关键技术，研制新型网络设备和软件系统，建设未来网络体系结构的

创新试验环境，研究内容中心网等各种新型网络体系结构和新协议，依托创新环境进行实验验证。

该 863 项目选择了 SDN 作为未来网络体系结构创新环境的设计思想。项目采用 SDN 思想提出了 FINE (Future Internet innovation Environment) 的网络体系结构，如图 33 所示。FINE 体系结构包括四个层次，即数据平面抽象和开放设备层、网域操作系统层、虚拟化平台层、新体系结构和新协议层。数据平面抽象和开放设备层对上层提供设备本地视图的编程接口供上层控制；网域操作系统层对上层提供全局的物理视图编程接口，方便上层使用（新体系结构或新协议可以直接编程，或通过虚拟化平台层来控制）；虚拟化平台层对上传提供特定的逻辑视图编程接口，为应用层提供虚拟化资源；还有一个比较重要的部分是域间协商通信机制，跨域不应该是一种控制机制，而应该是一种不同管理域的域间协商机制，通过网域操作系统的“东西向”接口，实现对邻居网络的通信链路、路由、资源和服务质量的协商。清华大学是国际上较早开展域间 SDN 研究的，提出了一种新型的协作式域间 SDN 机制 WE-Bridge，其研究开发成果在 INFOCOM2014 会议上进行了演示。

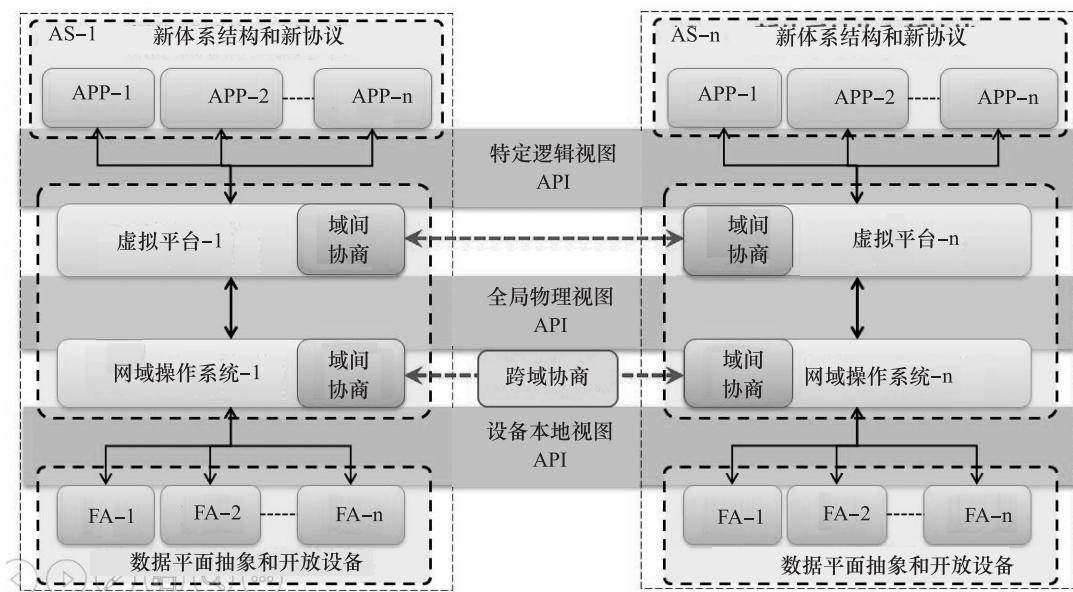


图 33 FINE 体系结构

2012 年，中美学术网络成立未来互联网工作组，由美国 Internet2 CTO Stephen Wolff 和清华大学毕军等担任共同主席，其工作组 Charter 的主要内容就是开展域间 SDN 的实验床研究，进行域间 SDN 的创新应用试验。工作组研究认为域间 SDN 实验床不能采用集中式的结构，因此采纳了清华大学所提出的 WE-Bridge 方案。在 WE-Bridge 设计和实现的基础上，在中美学术网络未来互联网工作组推动下，中国教育和科研计算机网 CERNET、美国 Internet2、中国科技网 CSTNET、荷兰学术网络 SURFnet 等合作，建立了首个基于协作式的跨洲际域间 SDN 实验床。该实验床从 2013 年 7 月开始运行，先后在 2013 年 9 月

中国杭州举行的 CANS 中美网络技术会议、2013 年 11 月美国 Denver 举行的 SuperComputing 国际超级计算会议、2014 年 1 月印尼万隆举行的 APAN 亚太先进网络会议、2014 年 3 月美国硅谷举办的 ONS 全球开放式网络峰会等国际会议上成功演示。Internet2 CTO Stephen Wolf 在 Internet2 官网上发表了对此技术的评价 (<https://www.internet2.edu/blogs/detail/5020/>)，认为目前 SDN 技术仅能用于单域环境，而此项开拓性工作展示了 SDN 技术可以扩展到全球科学合作之关键的多域环境，对美国部署 SDN 的校园网与 Internet2 国家基础设施互通也有重要意义。

浙江工商大学的 ForCES 课题组提出了基于 ForCES 的 SDN 体系结构，并将 SDN 体系结构分为应用层、控制层和基础资源层。浙江大学在 SDN 体系结构和网络操作系统两方面进行了研究，并在 2012 年 4 月第二届全球开放网络峰会上演示了基于 SDN 体系结构的可重构网络体系结构 XFlow。中科院计算所研制了构建 SDN 网络的可编程虚拟路由器 PEARL，该设备具备网络虚拟化功能并提供多种编程方法，可满足未来网络协议创新的需要。清华大学在 SDN 的系统架构以及 SDN 在数据中心网络中的应用方面开展了深入研究，并取得了创新性成果，如研究了以数据为中心的软件定义网络体系结构 SODA、基于 SDN 的数据中心网络虚拟化机制、基于 SDN 的数据中心网络组播协议、基于 SDN 的数据中心网络内容转发协议等，并在 SIGCOMM 2013 和 INFOCOM 2014 做了演示。解放军理工大学针对 SDN 的测量问题进行了研究，并研制了一种基于 OpenFlow 的未来互联网测量平台 OpenTrace。华为、中兴通信等也设立专门的研发团队，研发 SDN 交换机与控制系统，盛科网络研制了 OpenFlow 芯片。

解放军信息工程大学承担的“可重构信息通信基础网络体系研究”项目提出了“以变应变”的未来网络设计理念，以“强化基础互联传输能力”为突破口，创立了“网络元能力”的基本理论体系，构建可根据动态变化的特征要求和运行状态自主调整网络内在结构的关键机理和机制，力图在一张物理网上根据不同业务服务质量需求、不同时段业务量需求，重构出多个服务承载网，以解决目前的刚性网络结构难以适应业务类型不断增长的迫切需求。

同时，国内业界多次组织 SDN 的高峰会议和论坛。2013 中国计算机学会的年会上也组织了 SDN 论坛。SDN 技术和标准组织 ONF 于 2013 年与中国 SND 和开放式网络专委会签署战略合作协议，并授权了中国 BII 公司成立第一家美国以外的全球 SDN 认证测试中心。同时 2013 年华为、中兴、华三、神州数码等公司陆续推出了很多 SDN 技术产品。因此，2013 年在我国 SDN 发展历程上里程碑的意义。

3.2 面向服务的网络体系结构

在 973 项目“面向服务的未来互联网体系结构与机制研究”的支持下，中科院计算所提出了一种以服务层为“细腰”结构的互联网体系结构 SOFIA^[28]。SOFIA 在服务层实现服务的灵活处理，在网络层实现数据包高效转发，完成了服务灵活处理和数据高效传输的解耦合，使得二者可以分别发展，发挥各层优势。

我们可以从图 34 更加深入地理解 SOFIA 理念。从互联网实现的功能来看，主要包括图中所示的四个层次的内容：网络域、（网络中的）主机、（主机上的）服务以及（服务提供的）内容。IP 网络以主机地址（Host）为中心，通过划分不同的网络域（Domain）进行管理和互联。这种结构不需理解服务和内容，数据包处理速度高。但这也造成了数据冗余传输、安全性和移动性等问题。CCN/NDN 把设计的中心推向了另一个极端：以数据（Data）作为互联网体系结构的核心。这种面向内容的模型可解决移动性问题，提升信息分发效率，却由于过于极端支持内容分发，而对其他应用（如一对语音等）支持不够，即失去了灵活性。没有解决信息中心网络如何与现有互联网互联的问题。SOFIA 模型探索网络体系结构的设计空间，提出了介于内容和主机之间的新的抽象——服务。SOFIA 将网络中对数据的处理操作抽象称为服务，并用服务 ID 来唯一标识。相比于内容，服务具有更灵活的表达能力，而服务层可构建于现有的网络层之上，在数据传输的网络地址确定后，可利用网络地址转发数据，从而实现高效数据转发。

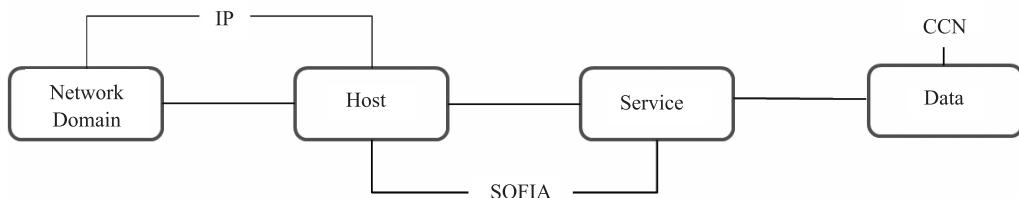


图 34 IP/SOFIA/CCN 模型原理

SOFIA 借助 SDN 思想，通过带外的控制器了解网络状况，并直接对服务层转发表进行管理，从而增强了网络侧的智能，提高了网络的可管性。目前，中科院计算所研究了网络内缓存、服务迁移等关键机制^[2]，实现了 SOFIA 协议栈原型，并研发了 SOFIA 网关，在 Android 系统上移植了 SOFIA 协议栈，开发了文件分发、移动语音通信等应用，这些应用已经在由几十个节点组成的小 CENI 平台上进行了部署和测试。

3.3 面向服务/内容的网络体系结构

北京交通大学承担的国家 973 项目“一体化可信网络与普适服务体系基础研究”创造性提出了以“两层模型、四种标识、三次映射”为典型特征的未来互联网新体系机理与架构，在网络体系中引入服务标识、连接标识、身份标识、以及交换路由标识，实现了身份标识与位置标识的分离和资源与位置的分离，综合有效解决了互联网在安全性、移动性、可扩展性何服务质量等方面的问题，研制了完整的原型系统并在多家单位推广应用，科技成果鉴定意见为“在未来信息网络体系、理论及技术等核心研究领域取得了重大突破性进展。在网络体系结构、标识解析映射机制等方面有重大创新，具有国际先进水平。”科技部结题评价为优秀，并获得 2012 年中国电子学会电子信息科学技术奖一等奖和 2013 年教育部技术发明奖一等奖。973 计划顾问组组长周光召先生曾带领部分

973 顾问组专家专程到北京交通大学考察，并给予高度评价。

在此基础上，该课题组进一步从网络与服务融合的角度，提出了智慧协协同网络体系结构，包含网络组件层、智慧服务层和资源适配层，并引入“实体域”和“行为域”，在“实体域”引入服务标识、族群标识与组件标识，在“行为域”引入服务行为描述、族群行为描述与组件行为描述，通过将资源与位置分离、身份与位置分离和控制与数据分离，力图综合有效解决互联网在可扩展性差、安全性差、移动性支持不足、资源利用率低、能耗高、用户体验差等方面的问题。

3.4 未来网络研究试验床

目前我国尚无国家级的未来网络研究实验床。但是，多个从事未来互联网相关研究的单位构建了各自的验证平台。例如，国家发展和改革委员会正在筹划构建 CENI (China Environment on Network Innovation) 平台；北京交通大学构建了一体化标识网络系统验证平台；解放军信息工程大学构建了可重构网络试验平台；清华大学和北京邮电大学也分别构建了 SDN 网络试验平台。

4 国内外研究进展比较

近几年，我国在互联网技术方面取得了一系列突破性进展，部分研究工作被国际相关著名期刊和学术会议接受，也被国际同行广泛认可。

在开放可编程网络方面，我国目前大部分研究工作处于跟踪水平。解放军信息工程大学提出的可重构信息通信基础网络体系，采取了“以变应变”的未来网络设计理念，以“强化基础互联传输能力”为突破口，创立了“网络元能力”的基本理论体系，可以说是这方面的一大亮点。

在面向服务的网络体系结构和内容中心网络体系结构方面，国内外的研究者从不同的角度，提出了若干未来互联网体系结构。总体而言，在设计思想和采用的技术路线方面，国内外基本处于同一水平。然而，由于互联网传统上被欧美垄断，而我国处于弱势地位，未来互联网研究的热点和重点仍然由欧美主导。同时，我国的大多数研究项目主要由高等学校和科研院所承担，极少有企业参与，使得产学研没有有效相互转换和互动。另外，国内对未来互联网体系的研究集中在技术方面，极少涉及未来网络体系对互联网产业的商业运行模式、对社会生活的影响等方面的研究。另外，项目间的竞争性和独立性，也造成重复研发、研发工作不能成体系，研究深度不足以及实际应用能力较差等弊端。

在未来网络研究实验床方面，我国尚无国家级的未来网络研究实验床，以开展未来网络体系的大规模试验验证，从而加强我国所提未来网络体系的实际应用能力。

5 发展趋势与展望

互联网已经渗透到包括政治、经济、文化、教育、卫生等人类社会生活的方方面面，在人类生活中发挥着重要作用。然而，现有互联网存在安全性差、可扩展性差、移动性支持不足、资源利用率低、能耗高、用户体验不佳、网络管理复杂等缺陷。因此，世界各国均投入巨资，研究未来互联网体系结构，力图解决现有互联网各种不足的同时，抢占未来信息网络领域的制高点。

虽然业界提出了各种各样的未来互联网体系结构，但总体趋势是解决现有互联网存在的三个绑定（即身份与位置绑定、资源与位置绑定、控制与数据绑定）问题。例如，SDN 侧重解决实现控制与数据的分离问题；面向服务的网络体系结构侧重解决资源与位置绑定问题；而面向移动性的网络体系结构侧重解决互联网的身份与位置绑定问题。

然而，仅仅解决互联网存在的某一个绑定，难以综合有效解决现有互联网存在的各种严重不足。因此，如何在一个网络体系结构下巧妙地解决互联网的三个绑定问题，从本质上综合有效解决前述严重不足，将是未来互联网体系结构研究的发展方向。

6 总结

本报告以未来互联网体系结构为核心，系统介绍了国内外在未来互联网体系结构研究方面的进展，并对国内外的研究进展进行了比较和展望，希望能够对国内相关的研究人员有所启发。

参考文献

- [1] Global environment for network innovations[EB/OL]. <http://www.geni.net>.
- [2] National Institute Of Information(NICT). AKARI Project[EB/OL]. <http://akari-project.nict.go.jp>.
- [3] Open Signaling Working Group[EB/OL]. <http://comet.columbia.edu/opensig/>.
- [4] TENNENHOUSE D L, WETHERALL D J. Towards an Active Network Architecture[J]. ACM Computer Communication Review. 1996, 26(2) : 2-15.
- [5] TURNER J, TAYLORD. Diversifying the internet[A]. Proceedings of the IEEE Global Telecommunications Conference(GLOBECOM'05)[C]. 2005, pp: 755-760.
- [6] WANG W M, HALEPLIDIS E, OGAWA K, et al. ForCES LFB Library, Work in Progress[EB/OL]. <http://www.tools.ietf.org/html/draft-ietf-forces-lfb-lib/>.
- [7] Software-Defined Networking: The New Norm for Networks[S]. ONF White Paper, 2012.
- [8] Empowering the Service Economy with SLA-aware Infrastructures. European Union 7th Framework Program

- [EB/OL]. <http://sla-at-soi.eu>.
- [9] SCHULZRINNE H, SEETHARAMAN S, HILT V. NetSerV-Architecture of a Service-Virtualized Internet. NSF NeTS FIND Initiative[EB/OL]. <http://www.nets-find.net/Funded/Netserv.php>.
- [10] COMBO[EB/OL]. <http://www.ict-combo.eu>.
- [11] ZHANGL Named Data Networking(NDN) Project[S]. Research Proposal, 2010.
- [12] KOPONEN T. A Data-oriented(and beyond) Network Architecture[A]. SIGCOMM'07[C] 2007. 181-192.
- [13] MobilityFirst, A Robust and Trustworthy Architecture for The Future Internet[EB/OL]. <http://mobilityfirst.winlab.rutgers.edu/Index.html>.
- [14] MOSKOWITZ R, NIKANDER P. RFC 4423: Host Identity Protocol (HIP) Architecture [S]. Internet Request for Comments, 2006.
- [15] MITSUNOBU K, MASAHIRO I. LIN6: A New Approach to Mobility Support in IPv6[J]. Internetional Symposium on Wireless Personal Multimedia Communication, 2000, (455): 1079-1083.
- [16] ANAND A XIA. An Architecture for an Evolvable and Trustworthy Internet[S]. 2011.
- [17] GENI PlanetLab[EB/OL]. <http://www.planet-lab.org/>.
- [18] PURSUIT [Online]. Available: <http://www.fp7-pursuit.eu/PursuitWeb/>.
- [19] [Online]. Available: <http://www.sail-project.eu/>.
- [20] FP7 COMET project. [Online]. Available: <http://www.comet-project.org/>.
- [21] FP7 CONVERGENCE project. [Online]. Available: <http://www.ict-convergence.eu/>.
- [22] 张宏科, 苏伟. 新网络体系基础研究——一体化网络与普适服务[J]. 电子学报, 2007, 35(4): 593-598.
- [23] 董平, 秦雅娟, 张宏科. 支持普适服务的一体化网络研究[J]. 电子学报, 2007, 35(4): 599-606.
- [24] 杨冬, 周华春, 张宏科. 基于一体化网络的普适服务研究[J]. 电子学报, 2007, 35(4): 607-613.
- [25] 863 计划信息技术领域 2008 年度专题课题申请指南. 见网页: <http://program.most.gov.cn/htmledit/B4E72055-86D8-34F6-F108-A1B43A835AEE.html>.
- [26] 国家高技术研究发展计划(863 计划)信息技术领域“三网融合演进技术与系统研究”重大项目申请指南。见网页: <http://program.most.gov.cn/htmledit/635C939B-506A-52A3-5398-E4E75FA19B50.html>.
- [27] <http://www.most.gov.cnfggwzfwj/zfwj2012/201202/W020120210626443434599.doc>.
- [28] Qinghua Wu, Zhenyu Li, Jianer Zhou, Heng Jiang, Zhiyang Hu, Yunjie Liu, Gaogang Xie. SOFIA: Towards Service-oriented Information Centric Networking. IEEE Network, May 2014.

作者简介

罗洪斌 工学博士, 教授/博士生导师。2007 年 6 月加入北京交通大学电子与信息工程学院; 2008 年 12 月破格晋升副教授; 2011 年 11 月破格晋升教授。2009 年 9 月 ~ 2010 年 9 月在美国普度大学 (Purdue University) 做访问学者。2012 年入选教育部新世纪优秀人才支持计划。研究方向为宽带通信网络。目前主持国家 973 计划课题、863 计划课题、国家自然科学基金面上项目各 1 项。曾主持和参与国家 973 计划项目、国家 863 计划项目、国家自然科



学基金等项目十余项。以第一作者身份在《IEEE/ACM Transactions on Networking》、《IEEE Journal on Selected Areas in Communications》等本领域高水平期刊和国际会议发表论文近 50 篇。

胡宇翔 男，博士，现任中国人民解放军信息工程大学副教授。长期从事宽带信息网络理论研究和工程开发，近 5 年累计发表论文 20 余篇，其中 SCI、EI 检索 10 余篇，申请国家技术发明专利 8 项，先后参加了 5 项国家级科研项目，其中作为主要参研人员参与国家 973 计划课题 2 项、作为子项负责人参加完成国家 863 计划课题 2 项、主持国家自然科学基金课题 1 项，出版专著 2 部。



毕军 毕业于清华大学计算机系，获学士、硕士、博士学位。曾赴美留学，美国贝尔实验室博士后、研究员。现任清华大学网络科学与网络空间研究院网络体系结构和 IPv6 研究室主任、教授、博士生导师。主要从事新型互联网体系结构和协议的研究和教学工作。发表 SCI/EI 收录的学术论文百余篇，获国家发明专利授权 20 余项，颁布或获批 RFC 国际标准 4 项。入选教育部“新世纪优秀人才”，曾获得国家科技进步二等奖、教育部技术发明一等奖、中国通信学会科学技术一等奖和二等奖。国家 863 项目“未来网络体系结构和创新环境”首席专家。担任国际学术会议主席 10 余次，亚洲未来互联网学会共同主席，中美学术网未来互联网工作组共同主席，中国 SDN 专委会常务副主任，中国计算机学会杰出会员、学术工委委员、互联网专委会委员、互联网学术年会程序委员会副主席等。



李振宇 中国科学院计算技术研究所，副研究员。主要研究方向为互联网体系结构、互联网测量。



深度学习的研究进展与趋势

CCF 人工智能与模式识别专业委员会

封举富¹ 王立威¹ 胡占义²

¹北京大学信息科学技术学院智能科学系，机器感知与智能教育部重点实验室，北京

²中国科学院自动化所，北京

摘要

深度学习是一类基于分层架构的机器学习算法。深度学习从数据中学习不同抽象层度的特征表示。本文介绍国内外深度学习及其应用的研究进展，并讨论深度学习所面临的挑战及发展趋势。

关键词：深度学习，分层架构，表示学习，卷积神经网络，受限玻尔兹曼机，自编码器

Abstract

Deep learning is a set of algorithms in machine learning based on hierarchical architecture. Deep learning attempts to learn multiple levels of representation and abstraction from data. In this paper, we review recent advances in the area of deep learning and its applications, discuss the main challenges and the future development of deep learning.

Keywords: deep learning, hierarchical architecture, representation learning, convolutional neural network, restricted Boltzmann machine, autoencoder

1 引言

人工神经网络的研究可以追溯到 20 世纪 40 年代，1943 年，McCulloch 和 Pitts 提出了第一个神经元的数学模型^[1]。1949 年，Hebb 提出了神经元学习准则^[2]。1958 年，Rosenblatt 提出了感知机（perceptron）^[3]，是第一种可以通过样本学习来改变神经元连接权重的神经网络，开创了人工神经网络研究的第一次热潮。但是，它本质上还是一个线性分类器，因此好景不长。1969 年，Minsky 和 Papert 出版了《Perceptron》一书，指出单层感知机不能解决 XOR 问题，并且“我们直觉判断推广到多层系统也不会有好结果，但是对于这一点我们认为证明（或否定）它是一个很重要的需要研究的问题。”^[4]由于 Minsky 在人工智能领域的巨大影响力，人工神经网络研究随之进入了长达 10 多年的低潮期。事实上，多层感知机是可以求解线性不可分问题的，但在当时缺乏有效的算法。实际上，1963 年在解决某些控制问题时，反向传播算法（Backpropagation，BP）已经被提

出^[5]，随后又在 1970 年^[6]和 1974 年被先后发现^[7]，但都不为人工神经网络研究界所知。直到 1986 年，反向传播算法才被重新发现^[8]，由此进入人工神经网络研究的第二次热潮。其中一个重要的研究成果是理论上证明了只含一个隐层的前馈网络可以在闭区间上一致逼近任意连续函数^[9,10]。1991 年，Hochreiter^[11]发现传统的 BP 算法用在深度网络的训练时，会产生所谓的“Long time lag problem”，即累积传递误差会减小的太快，或增加的太快（either shrink rapidly or grow out of bounds），因而无法对深度网络进行有效训练。尽管 1989 年 LeCun 等人利用卷积神经网络（Convolution Neural Networks, CNN）在手写数字识别中取得当时世界最好结果^[12]，但其在大规模自然图像上表现不佳。因此，实际应用中人们普遍使用浅层神经网络结构。在人工神经网络设计和训练过程中往往需要更多的经验和技巧，参数的确定和容易陷入局部极值等成为广受诟病的问题。Vapnik^[13]更是提出了严厉的批评：“虽然在一些特殊领域中应用神经网络取得了很重要的成果，但是所取得的理论成果并没有对一般的学习理论带来多大贡献。而且，在神经网络的实验中也没有发现新的有意义的学习现象。”随着支持向量机（Support Vector Machine, SVM）^[14]的出现，统计学习成为机器学习研究中的热潮。支持向量机最初就叫 SupportVector Networks，表明它本质上是一类特殊的浅层神经网络，但是其目标函数是凸的，因此具有全局最优解。SVM 可以说是万千宠爱集一身，漂亮的理论、优异的性质、高效的学习算法，以及在数字识别中取得了更好的结果^[15]，使得它成为新的研究热点，很多神经网络的研究者都纷纷转向研究 SVM。人工神经网络研究遭遇了又一次寒流。

随着互联网的迅速发展和大数据的出现，给机器学习带来了新的机遇和挑战。对于机器学习而言，大数据之“大”一方面体现在特征维数非常高，如图像、语音和文本通常是成千上万维，另一方面数据量非常大，通常是百万、千万甚至上亿量级的数据，并且一般缺少标签、不够精确。由于样本采集的困难和计算能力的限制，无论是神经网络还是统计学习算法处理的往往是小样本、低维数据问题。小样本、高维数据会带来所谓“维数灾难”问题^[16]，各种算法容易出现过拟合的现象，其中的关键问题是高维函数远比低维函数复杂。人们在分析大量实际问题中发现，高维数据实际上嵌入在一个低维的流形上。因此，特征选择和提取成为一个关键问题。Viola 和 Jones 利用级联 AdaBoost 从 45396 个过完备的 Harr 特征中进行多层特征选择，实现了人脸的实时检测^[17]。2000 年发表在《Science》上的两篇论文^[18,19]引领了流形学习的潮流，其本质在于通过非线性变换去除高维特征之间的相关性，从而揭示高维空间中数据分布的内在低维属性，即流形的内蕴几何结构。对于传统机器学习而言，初始特征往往是领域专家预先设计好的，比如，对尺度、旋转以及一定视角和光照等图像变化都具有不变性的 SIFT 特征（Scale Invariant Feature Transform）^[20]，在计算机视觉领域得到广泛应用。特征选择和流形学习不过是从中选择出“好”的特征或者去除高维特征之间的相关性。

2006 年，Hinton 等人^[21]在《Science》上发表的论文开启了深度学习的浪潮，也可以说是人工神经网络研究的第三次热潮。其实 Hinton 等人提出的深度信念网络从结构上看与传统的多层感知机区别不大。主要不同体现在做有监督学习前要先做预训练（pretraining）——非监督学习，并将预训练得到的权值作为有监督学习的初值进行微调

(finetune) 训练。预训练采用的是受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM)^[22,23]。一个 RBM 预训练好后固定权值，其隐层作为下一个 RBM 的输入层，多个 RBM 叠加起来就构成一个深度网络的编码器，编码器倒过来就是一个解码器，最后进行有监督学习的微调训练（图 1）。在 MINIST 手写数字识别实验中，用的是 784-500-500-2000-10 的网络结构，其识别误差率为 1.2%。

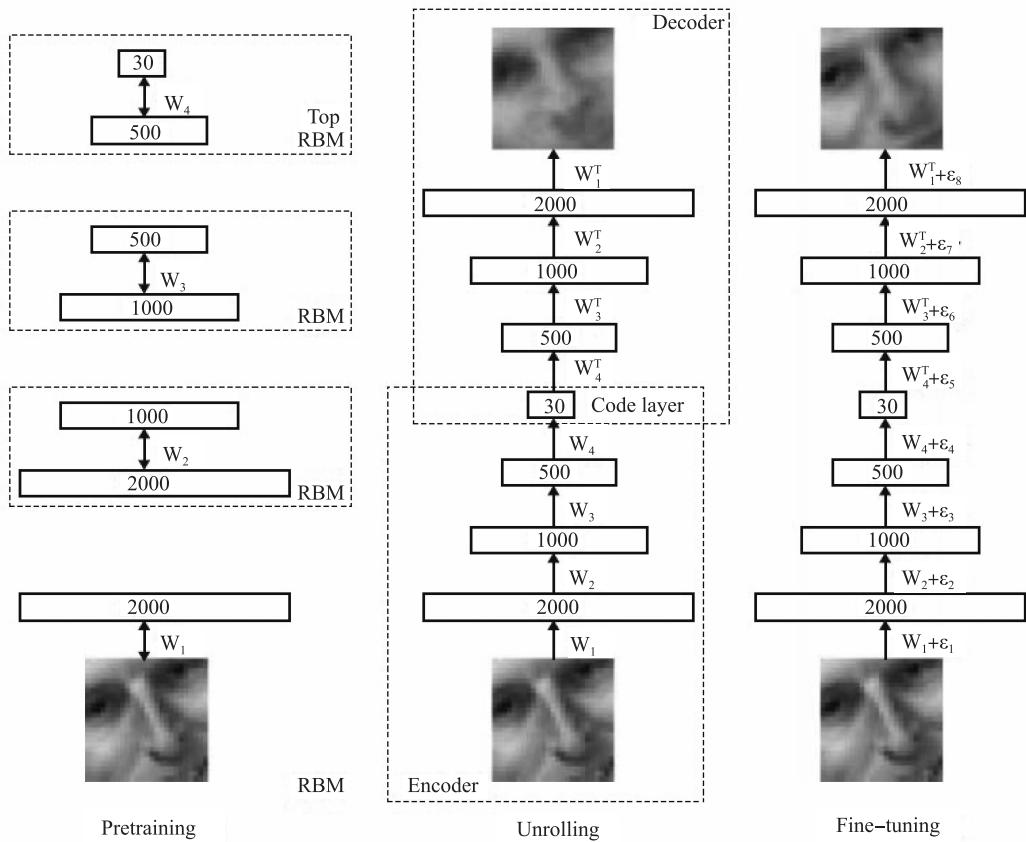


图 1 叠层 RBM 的预训练和微调^[21]

这项工作说明，深度神经网络可以通过自学习得到优异的特征，对数据有更本质的刻画，有利于可视化或分类；复杂的深度神经网络可以通过逐层 (layer wise) 预训练来有效克服训练的困难。LeCun 在 2014 年 2 月 10 日的一次访谈中说^[24]，2003 年他就和 Hinton、Bengio 等人将研究重点放在表示学习 (Representation Learning) 这个问题上。他和 Vapnik 也经常讨论深度网络和核函数的优缺点。实际上，SVM “不过是简单的两层模型，第一层是用核函数来计算输入数据和支持向量之间相似度的单元集合。第二层则是线性组合了这些相似度。”“第一层就是用最简单的无监督模型训练的，即将训练数据作为原型单元存储起来。”LeCun “评价核方法是一种包装美化过的模板匹配”，“‘窄’核函数所产生的支持向量机，通常在训练数据上表现非常好，但是其普适性则由核函数的宽度以及对偶系数决定。Vapnik 对自己得出的结果非常自信。他担心神经网络没有类似

这样简单的方式来进行扩展控制（虽然神经网络根本没有普适性的限制，因为它们都是无限的 VC 维）。我反驳了他，相比用有限计算能力来计算高复杂度函数这种能力，扩展控制只能排第二。在进行图像识别时，移位、缩放、旋转、光线条件和背景噪声等问题，会导致以像素做特征的核函数非常低效。但是对于深度架构（比如卷积网络）来说却是小菜一碟。”

Hinton 等人的工作极大地鼓舞了人工神经网络研究人员的士气，吸引了一大批优秀的学者加入，获得了学术界和工业界的高度关注和广泛重视。2012 年 6 月，《纽约时报》披露了 Google Brain 项目，由 Andrew Ng 和 Jeff Dean 共同主导，利用 16000 个 CPU Core 的并行计算平台来训练一种内部含有 10 亿个节点的“深度神经网络”模型，在语音识别和图像识别等领域获得了巨大的成功。2012 年 11 月，微软在天津展示了智能同声传译项目，讲演者用英文演讲，后台的计算机自动完成语音识别、机器翻译和中文语音合成，效果非常流畅。2013 年 1 月，百度宣布成立深度学习研究院（IDL）。2013 年 3 月，Google 收购了 Hinton 与他的两个研究生 Alex Krizhevsky 和 Ilya Sutskever 的创业公司 DNN Research。2013 年 12 月，Facebook 在纽约创建了深度学习人工智能实验室，Yann LeCun 兼任该实验室主管。2014 年 1 月，谷歌以 4 亿美金收购深度学习创业公司 DeepMind Technologies。2014 年 5 月 17 日，吴恩达（Andrew Ng）正式加盟百度，担任百度首席科学家，全面负责百度研究院。2014 年 7 月 14 日，微软展示了 Adam 项目，其目标是让计算机能识别任何物体。该系统建立了一个由 20 亿个连接组成的神经网络，使用的机器数量为同类系统的 1/30。

2014 年 4 月，《MIT Technology Review》将深度学习列为 2013 年十大突破性技术（breakthrough technology）之首^[25]。目前，深度学习在图像分类、语音识别、自然语言处理等方面取得了巨大的成功，已成为互联网大数据和人工智能的一个新的研究热潮。

2 国际研究现状

机器学习框架大体包括输入、特征提取、特征选择以及分类或预测等部分（图 2），传统机器学习研究主要关注分类或预测模型的设计（部分涉及特征选择），而把特征提取的任务留给领域专家。

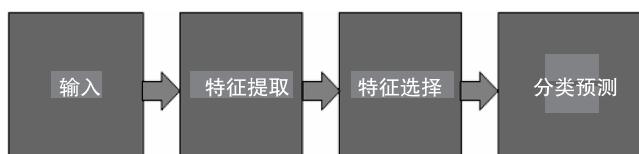


图 2 机器学习框架

与传统的机器学习方法不同，深度学习充分利用大数据的特点，自动学习不同抽象层度的特征表示，进而提高分类和预测的准确性。Bengio 认为深度学习最重要的目的就

是学习一个好的数据表征^[26]。深度学习利用包含多个隐层的人工神经网络来进行学习，隐含节点对应了输入信号变换后的特征，并且是逐层抽象的，所学到的特征对数据有更本质的刻画^[27]。如图3所示^[28]，网络的输入层是图像像素，第一个隐层学到的是带有方向性的边缘，对应视觉的底层特征，后面的隐层是前一层特征的组合，对应视觉的中层特征，最高层对应语义特征。

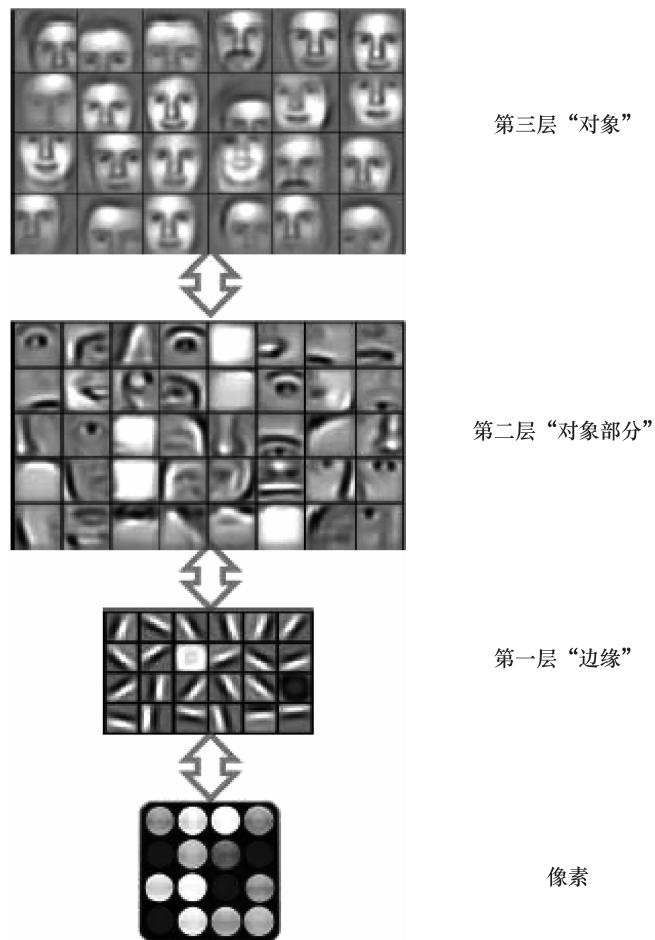


图3 学习特征分层 (Learning Feature Hierarchy)^[28]

深度学习是一类基于神经网络的机器学习算法，网络结构包含两个以上非线性隐含层。目前，一些常用的网络结构包括卷积神经网络、深度信念网络（Deep Belief Network, DBN）、深度玻尔兹曼机（Deep Boltzmann Machine, DBM）和自编码器（Autoencoder, AE）等。

2.1 卷积神经网络

卷积神经网络（CNN）是第一个成功训练多层网络结构的学习算法。它具有共享网

络权值的特点，降低了模型复杂度。在卷积网络中图像的局部区域作为最底层的数据输入，每层通过一个滤波器来获取特征，对图像的平移、缩放、倾斜等具有不变性。对于每层不同的神经元，他们共享对应的滤波器，从对每个神经元训练一个特定的滤波器变为对同一层的每个神经元训练多个共同的滤波器，通过这种权值共享减少了网络自由参数的个数。

LeNet5 是一种典型的用来识别数字的神经网络，由 Y. LeCun 提出^[29]，被广泛应用在许多银行识别支票上手写数字的商业用途中，结构如图 4 所示。其中卷积的那些操作即是对应的滤波器，子采样（Subsampling）的操作对应图像的局部采样过程。CNN 在语音识别、文档分析等也有广泛的应用。

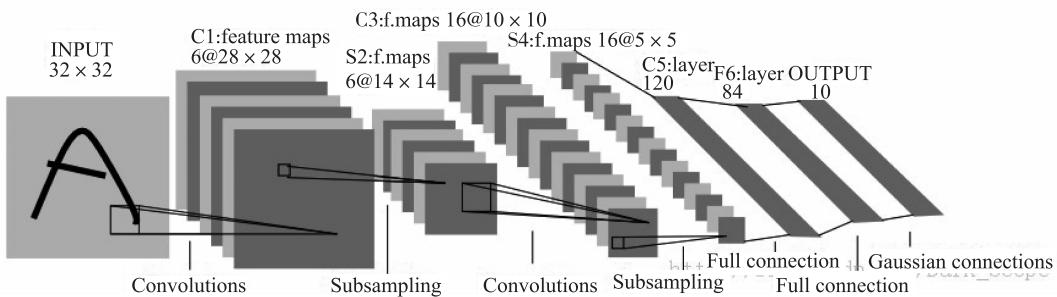


图 4 CNN 的典型示例 LeNet-5

2.2 受限玻尔兹曼机

受限玻尔兹曼机（RBM）是一个概率模型，由显层（ v ）和隐层（ h ）组成，显层与隐层之间是全连接的，但层内神经元之间没有连接。各层的节点都是一些 0-1 二值的随机变量（输入为连续值的可由高斯分布处理）， $P(v, h)$ 服从玻尔兹曼（Boltzmann）分布， $\theta(w, a, b)$ 是参数。

$$E(v, h; \theta) = -v^T w h - b^T v - a^T h$$

$$P_\theta(v, h) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta)); \quad Z(\theta) = \sum_{v, h} \exp(-E(v, h; \theta))$$

$$P_\theta(v) = \frac{1}{Z(\theta)} \sum_h \exp(-E(v, h; \theta))$$

给定独立同分布训练样本 $D = \{v^1, v^2, \dots, v^N\}$ ，由最大似然估计可以得到，

$$L(\theta) = \frac{1}{N} \sum_{k=1}^N \log \left(\frac{1}{Z(\theta)} \sum_h \exp(-E(v^k, h; \theta)) \right)$$

$$\frac{\partial L(\theta)}{\partial w_{ij}} = \frac{1}{N} \frac{\partial}{\partial w_{ij}} \sum_{k=1}^N \log \left(\sum_h \exp(-E(v^k, h; \theta)) \right) \frac{\partial}{\partial w_{ij}} Z(\theta)$$

$$= E_{\text{data}}[v, h] - E_{\text{model}}[v, h]$$

前后两项分别代表样本期望和模型期望。其中模型期望的计算较为复杂，Hinton 使用了

一种 Contrastive Divergence 的方法来代替 Gibbs Sampling，这样在避免超大计算量的同时仍有较好的近似^[30]。Roux 和 Bengio 证明了 RBM 可以一致逼近任意离散分布^[31]。对于具有高维特征结构的数据，RBM 可以很好地提取其中隐含的特征（图 5）。RBM 有多种变化，如在 RBM 的对数似然函数中加入稀疏惩罚项，则可以得到稀疏 RBM^[32]，还有 spike-and-slab RBM^[33]。

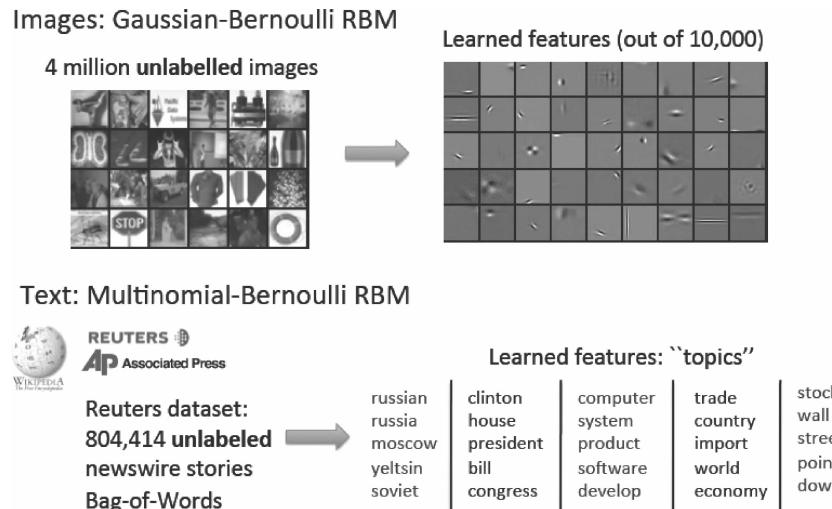


图 5 RBM 提取特征^[34]

如果把 RBM 的隐层数增加，可以得到 DBM^[35]；如果再将靠近数据层的部分层之间的链接变成有向图就可得到 DBN^[21]（图 6）。

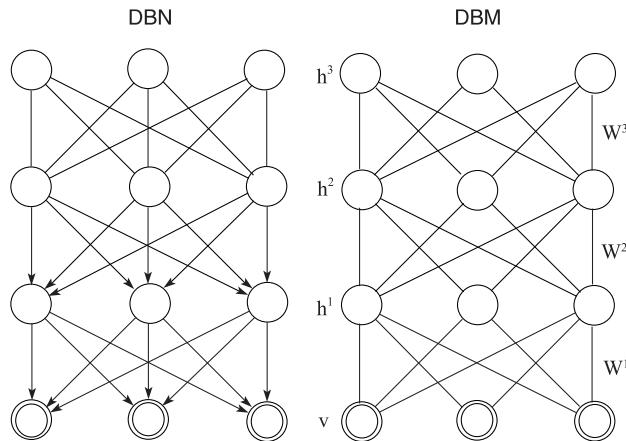


图 6 DBN 和 DBM^[35]

DBN 的训练是先自底向上进行逐层预训练。预训练每一层时将低的层看成是显层，高的层看成是隐层，再对整个模型进行微调。这种贪心的训练方法被验证是高效的。

DBM 的训练更加复杂些，在预训练隐层时需要考虑前一个隐层（自底向上）以及后一个隐层（自顶向下）的依赖关系。Salakhutdinov 和 Hinton 针对预训练给出了一个变化

性的下界，并据此设计了一个更好的预训练算法^[36]。Goodfellow 等人则提出了一种不需要预训练的 DBM 训练方法^[37]。该方法利用 DBM 的平均场（Mean-Field）方程训练一个可以进行多种推理任务（在文章中主要是预测部分变量）的递归网络（recurrent networks）。

2.3 自编码器

第一个利用无标签数据进行预训练的思想可能是 1987 年 Ballard 的工作^[38]。自编码器（AE）利用无标签数据自动提取特征进行编码，通过解码重构的数据与原数据比较来进行预训练，然后逐层（Layer-wise）训练编码器得到多层特征，每一层的输入都是上一个编码器的输出，最后进行有监督学习的微调训练（图 7）

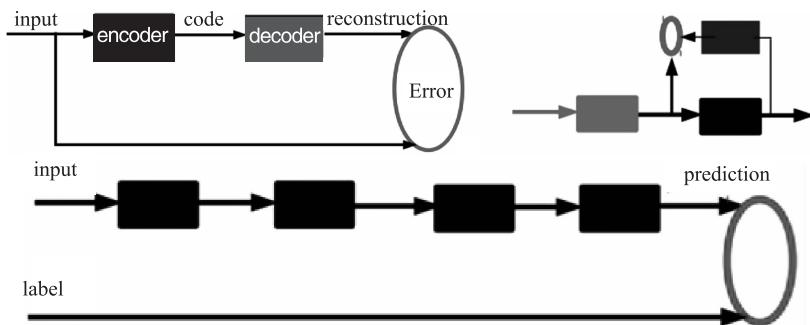


图 7 自编码器^[39]

给定输入，在特征表示中加入稀疏惩罚项，则有如下目标函数：

$$L(x; w) = \|wh - x\|^2 + \lambda \sum_j |h_j|$$

在进行无监督学习过程中，Y. Bengio 等人通过在数据层加入随机噪声来提取鲁棒特征，提出了去噪自编码器（Denoise Autoencoder, DAE）^[40]。DAE 将样本的某些输入（通常是约一半）设为 0，这个过程使得算法可以通过部分输入信息来估计缺失信息，以达到去噪的目的（图 8）。

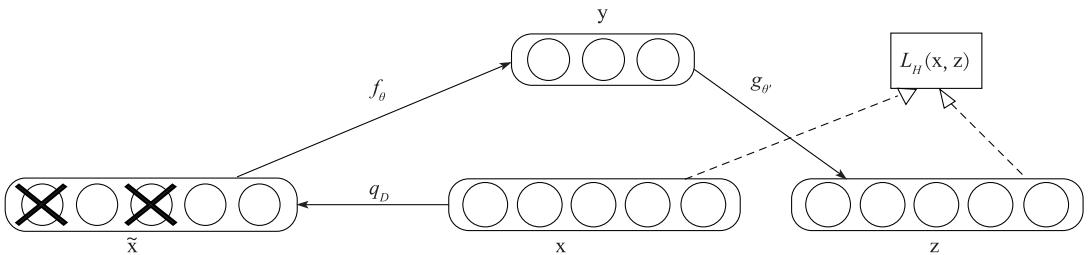


图 8 去噪自编码器^[40]

从信息论的观点看，最小化重构误差等价于最小化污染部分数据和原数据的某种相关信息量的下界，因此这种方法可以从输入的污染数据找到与它包含的信息量接近相同

的完整数据，进而达到去噪的目的。

在训练方法方面，为避免过拟合，Hinton 等人提出了一个简单高效的 Dropout 方法，在训练时每个神经元以一定的概率不工作^[41]，在 MNIST、TIMIT、CIFAR-10 等数据集上取得了很好的效果^[42]。Ba 和 Frey 提出了一个自适应的 Dropout 方法，与普通的 Dropout 方法不同，每个神经元不工作的概率并非是固定的，而是通过一个神经网络学习出来的^[43]。Wan 等人提出了一个类似的 Dropconnect 方法，每对神经元之间的相互作用以一定的概率不工作^[44]。Wang 和 Manning 提出了快速 Dropout^[45]。Baldi 和 Sadowski 指出^[46]，Dropout 与模型的归一化加权几何平均存在某种关联，而模型的归一化加权几何平均是对模型的期望的一个近似，因此可以认为 Dropout 与模型平均（Model Averaging）具有类似的效果。Dropout 还可以看做是对梯度的一种正则化，Wager 等人揭示了 Dropout 与自适应 L2 正则化、AdaGrad 之间的联系^[47]，并进一步证明一定条件下 Dropout 可以指数级地改善 ERM 泛化错误率的界^[48]。另外，MCF（Marginalized Corrupted Features）算法^[49]、mDAE（marginalized Denoising Auto-Encoder）^[50]、Maxout 网络^[51]也取得了很好的结果。

在大规模的学习方面，Google 的 Jeffrey Dean 等人提出了一个基于 CPU 的并行异步框架 DistBelief 来训练深度神经网络^[52]，包括两个并行算法：1) 异步随机梯度下降方法，即将数据划分成若干部分，每部分由一个神经网络进行训练，这些神经网络周期性地同步参数；2) 分布式 L-BFGS 算法，即分布式地将模型参数存储在多个服务器上，使用一个协调进程来指挥各个服务器进行独立的运算。该框架可以调用上万个 CPU 核来训练有 10 亿个参数的神经网络。很多研究者采用 GPU 加速训练过程^[53~56]。Andrew Ng 团队以一个廉价的 GPU 集群成功进行了猫脸识别实验^[57]。

目前，深度学习在理论方面进展十分有限。原因之一是对于深度神经网络这样一种非凸模型，为什么能够被充分训练并在实践中取得非常好的效果还缺乏认识。2014 年，著名理论计算机科学家，2 届歌德尔奖得主 S. Arora 教授等人提出了一个有严格理论依据的算法来学习生成式的神经网络模型^[58]。他们设计了一个逐层学习的算法，并证明对于一类由随机的比较稀疏的深度神经网络生成的 ground truth，该算法可以以大概率在多项式时间内，利用不超过 3 次多项式个样本，输出一个与真实分布接近的模型。尽管此理论结果有一定意义，但该工作对于深度神经网络结构做了较多的假设，与实际所用的网络有一定差异，因此该理论的意义还有待检验。

尽管深度学习在理论方面进展有限，但在应用方面却取得了巨大的成功。2011 年，在 IJCNN 2011 交通标志识别竞赛（Traffic Sign Recognition Contest）中，基于 CNN 的方法首次取得了比人的识别率还要高的结果^[59]。2011 年，微软基于深度神经网络的语音识别彻底改变了语音识别原有的技术框架^[60]。2012 年，Hinton 领导的研究小组使用了深度学习技术，在著名的 ImageNet 问题上取得世界最好结果^[42]。2012 年，Andrew Ng 团队利用 1 000 台机器共 16 000 个核训练了一个含有 10 亿个参数的网络，在 20 000 多个类别的 ImageNet 问题上取得世界最好结果^[61]。2013 年，基于 DNN 的方法将汉字仅仅看做“图形”，取得了当时最高的手写汉字识别率^[62]。基于长短时记忆

(Long Short Term Memory, LSTM) 的声学建模和递归神经网络的方法打破了 TIMIT 语音(音素)的识别记录^[63]。近年来, 基于深度学习的算法可以说捷报频传, 先后在 NIST OpenHaRT2013^[64]、TheMICCAI 2013 Grand Challenge on Mitosis Detection^[65]、PASCAL object detection^[66]等取得了最好成绩。另外, 深度学习在场景标识^[67]、目标检测^[68]、阴影检测(shadow detection)^[69]、视频分类^[70]、卫星图片标记^[71]、蛋白质结构预测^[72]、语言模型^[73-75]等获得了成功应用。

3 国内研究进展

国内关于深度学习的研究在学术界和工业界均有开展。2013 年 1 月, 百度宣布成立深度学习研究院(IDL)。百度研究院常务副院长余凯博士领导的团队率先开展深度学习方面的研究和成果转化^[76], 目前已经有超过 8 项深度学习技术在百度产品上线, 尤其在稀疏编码方面取得了公认的成绩^[77,78]。他们设计了一个三维 CNN 来识别人体动作^[79], 在 TRECVID 和 KTH 上都取得了很好的实验效果。华为诺亚方舟实验室的研究人员设计了一个新的深度神经网络结构来学习短句的匹配(如问答系统)^[80]。基于自然语言的局部性与层次性, 该方法首先使用 LDA 建立带有分层结构的主题模型, 得到词汇共同出现的模式; 然后建立一个稀疏的深度神经网络, 在网络的最高层使用 Logistic 回归得到一个匹配的分数。微软的研究人员提出了一个上下文相关的结合深度神经网络与隐马尔可夫模型(HMM)的模型(CD-DNN-HMM), 用于语音识别^[81], 在 SWBD-I 数据集上取得了很好的效果。

香港中文大学汤晓鸥和王晓刚团队从 2011 年开始开展深度学习研究, 将深度学习模型应用于人脸识别、行人检测、姿态估计、人体图像分割、车型识别、大规模人群监控、通用物体识别和检测、互联网图像检索等^[82-93], 取得了一系列先进成果, 尤其是他们研发的 DeepID 人脸识别技术在 LFW (Labeled Faces in the Wild) 数据库上获得了 99.15% 的识别率, 而如果仅仅给出人脸中心区域, 人用肉眼在 LFW 上的识别率为 97.52%。Facebook 发布的人脸识别算法 DeepFace, 在 LFW 上的识别率为 97.35%。DeepFace 需要 700 多万人脸数据作为训练, 而 DeepID 仅使用了 20 万张人脸数据以及数台 Nvidia K40 GPU。最近他们还设计了一个深度感知器, 能够在输入单个二维脸部图像时学习出它各个不同角度下的脸部图像, 用以模拟大脑在二维图像中感知三维图像的能力, 在 MultiPIE 数据集上取得了很好的结果^[94]。香港城市大学的团队设计了一个基于卷积神经网络的多任务模型 HMLPE, 应用于人体姿态估计, 在 Buffy Stickmen、ETHZ Stickmen 等数据集上取得了有竞争力的结果^[95]。

清华大学张长水研究组提出了一种改进的卷积神经网络训练方法, 在交通标志识别数据集上取得了 99.65% 的最高识别率^[96]。针对音乐数据构建了深度信念网络和级联自编码器的混合模型, 在古典作曲家分类任务上取得了 76.26% 的最高识别率^[97]。清华大学孙茂松研究组提出了基于递归自动编码器(Recursive Autoencoder)的调序模型, 通过

计算变长字符串分布式表示缓解了数据稀疏和特征设计问题，提高了基于反向转录文法和基于短语的翻译系统的性能^[98,99]。清华大学胡晓林等人提出给 HMAX 模型加上稀疏性约束，用稀疏学习得到的模版代替随机选取的模版，使得 HMAX 在图像分类任务上性能得到大幅提升^[100]。他们构建的一个逆分层模型在预测图像显著性方面也得到了不错的结果^[101]。清华大学张钹和朱军研究组提出了 Dropout-SVM^[102]。中国科学技术大学陈恩红研究组提出了一种稀疏自编码器来进行图像去噪^[103]，取得很好的结果。北京大学吴玺宏研究组实现了不需要事先对序列数据切分的 DNN-HMM 模型训练^[104]，提出了基于 DNN 得到声学状态向量表示的决策树聚类算法^[105]，从而实现了在没有 GMM 的基础上得到上下文相关的建模单元，通过引入词性信息在一定程度上解决了汉语中一词多义和同形异义的问题^[106]。北京大学谢昆清研究组提出了基于异质多任务深度学习的交通流量预测方法^[107]。哈尔滨工业大学和微软研究院合作的研究团队提出了图像检索的深度神经网络模型，提升了 Bing 的图像检索效果^[108]。清华大学和微软研究院合作的研究团队设计了词表示学习的一种深度模型，在 WordSim-353 等数据集上大幅提高了词表示学习的有效性^[109]。

中科院自动化所刘成林和向世明研究团队提出了一种可嵌入多尺度卷积核的深层卷积神经网络模型，提高了网络对不同尺度的视觉目标的描述能力；随后他们提出了一种平行深层卷积神经网络模型，通过对卷积层和 max-pooling 层的平行结构化分离以及对其输出的隐层融合，增强了网络对不同类型图像数据的可学习能力，在基于遥感图像的城市车辆目标检测的应用中验证了该模型的实用性^[110,111]。中科院自动化所谭铁牛研究团队提出了一种多任务深度神经网络模型，通过为每个类别标签附加“正”和“负”节点来扩展网络的多标签学习性能，并在自然图像标注中取得了较高的标注精度^[112]。中科院自动化所李子青研究团队使用双卷积神经网络学习了非线性度量，在 VIPeR、PRID 2011 等数据库上取得了很好的效果；同时，他们研究了跨库（Cross Dataset）条件下的行人再识别问题，验证了该技术的实用性^[113]。中科院计算所陈熙霖研究组提出了一种耦合神经网络 DCAN (Deeply Coupled Autoencoder Networks) 用以处理图像的多视角分类问题，在 MultiPIE 等数据集上取得了很好的结果^[114]。中科院声学所颜永红研究组提出了一种利用循环式神经网络语言模型（Recurrent Neural Network-based Language Models, RNNLM）在语音识别词网上重打分的算法^[115]。

在深度学习的理论方面，南京大学周志华教授的研究团队为 Dropout 在神经网络中的作用给出了理论分析，他们对多种类型的 Dropout（对隐藏元、输入元或者对权重的 Dropout）的 Rademacher complexity 进行了分析，证明了 Dropout 使得单层神经网络的 Rademacher complexity 有多项式级的下降，而在深度神经网络下有指数级的下降^[116]。

目前，国际上深度学习的研究发展迅猛。尽管国内研究开展时间不长，与国际先进水平还有差距，但已在个别方向取得突破性进展，已经引起广泛关注和重视。2013 年 7 月，由人工智能与模式识别专业委员会协办的中国计算机学会人工智能会议特别邀请了百度深度学习研究院常务副院长余凯博士做特邀报告，并为会议专辑撰写了论文“深度学习的昨天、今天和明天”^[117]。2013 年 11 月，在第十一届中国机器学习及其应用研讨

会上，百度深度学习研究院首席科学家张潼做了“大数据和深度机器学习介绍”的报告。相信经过广大研究人员的努力有望赶超国际先进水平。

4 发展趋势与展望

深度学习已成为当前的热点，并有望引领未来的研究发展方向。表面上看，似乎与 20 多年前的多层神经网络区别不大。理论上只含一个隐层的前馈网络可以在闭区间上一致逼近任意连续函数。也就是说，对于任意非线性函数，浅层网络和深度网络都能得到足够好的表示。但是深度模型可以将复杂函数分解为简单函数的逐层组合，这样所需参数以及训练样本就少。尽管深度学习在许多应用领域取得了巨大成功，但是如何构建一套坚实的理论基础仍然任重道远。

分层网络结构和对应的学习算法是深度学习的核心问题。结构决定潜力，学习算法在于挖掘这种潜力。人类视觉的物体识别过程主要由腹部通道完成 (Ventral pathway)。从 V1 区到 V2 区的简单特征处理，到 V4 区、PIT 区和腹部通道最高级别的 AIT 区的高层处理，腹部通道的物体识别过程是一个由简单到复杂的分层加工过程。我们觉得，深度神经网络在物体表达和识别方面的巨大成功，可能正是从某种侧面揭示和利用了这种分层架构 (hierarchical architecture) 固有的潜力。2012 年，MIT 的 Poggio 在韩国召开的亚洲计算机视觉的特邀报告上曾说，十多年来一直困惑他的一个问题是他提出的图像物体识别模型 Hmax，一种简单的分层模型为什么取得如此好的结果？2013 年，他们推测，这种分层模型在于学习了“由简单到复杂的不变量”，并给出了系统的理论分析^[118]。当前，场景理解中广泛使用的分层条件随机场优化框架 (Hierarchical Conditional Random Field)^[119]，从像素层→超像素层→特征层→部件层→物体层→场景层，也体现出巨大的潜力和优势。George 和 Hawkins 基于神经生理提出的物体表达和识别模型 HTM (Hierarchical Temporal Memory) 也体现出巨大的潜力和优势。所以，“深度分层结构”可能是问题的本质，同时“隔层输入”（而不是目前主流方法的逐层输入）和“高层反馈”网络结构构建和对应的学习算法可能是今后一项需要深入研究的内容。

在具体实现方面，我们需要多少训练样本才能学习到足够好的深度模型？事实上，对于给定的训练样本集，要找到精确匹配的网络模型是 NP 难问题^[121~123]。由于深度模型都是非凸函数，是否有更好的优化算法^[124]？针对具体应用问题，如何设计适合的深度模型？比如，涉及结构化信息的自然语言处理如何建模？如何确定网络层数、每层节点数、初始化参数等广受诟病的问题，在深度网络中依然存在，需要更多的经验和技巧^[125]。另外，现有训练方法大都是采用随机梯度法，无法在多个计算机之间并行。即使采用 GPU 其训练时间也非常漫长，比如训练几千小时的声学模型可能需要几个月时间。因此，研发适合深度网络快速学习的软硬件和优化算法将是其中的一个重要课题。

展望未来，我们是否离所谓的“奇点”很近了？Ray Kurzweil 预言 2029 年机器会通

过图灵测试^[126]。LeCun 说“人工智能的每一个新浪潮，都会带来这么一段从盲目乐观到不理智，最后到沮丧的阶段。”那么，这次是否真的不一样？人工智能之父 Marvin Minsky 说“What magical trick makes us intelligent? The trick is that there is no trick. The power of intelligence stems from our vast diversity, not from single, perfect principle.”

5 结束语

总之，大数据的出现给机器学习带来了新的机遇和挑战。深度学习可以充分利用大数据的特点，自动学习不同抽象层度的特征表示，突破了传统机器学习的瓶颈，可望揭示这种“多样化”的信息处理机制。目前，国际上正在兴起的模拟脑计划，如美国 2013 年启动的为期 10 年耗资 10 亿美元的“BRAIN”计划，欧共体 2013 年启动的 flagship 为期 10 年耗资 10 亿欧元的 Human Brain Project，将为深度网络提供更多的脑网络结构和认知基础，同时，深度学习也将为揭示脑信息加工机理提供计算和模拟手段。我们认为随着研究的深入，深度学习将引领未来的发展方向，促进神经科学与计算机科学的交叉融合，并有望加速推进人工智能向前发展^[127]，可望成为智能信息处理的一种颠覆性技术，其意义是深远的。

致谢

本文授中国计算机学会人工智能与模式识别专委会委托撰写。作者感谢南京大学周志华教授、北京交通大学于剑教授、清华大学张长水教授、北京大学吴玺宏教授和谢昆清教授的指导和帮助，并感谢中国计算机学会学术工委提供的建议和参考文献。限于时间和水平，难免遗漏了许多国内外重要工作，我们表示歉意。本文得到国家自然科学重点基金（61333015）和国家重点基础研究发展计划（2011CB302400）资助。

参考文献

- [1] McCulloch W, Pitts W. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics*, 1943, 7: 115-133.
- [2] Hebb D. *The Organization of Behavior* [M]. Wiley, New York, 1949.
- [3] Rosenblatt F. The Perceptron: a Probabilistic Model for Information Storage and Organization in the Brain [J]. *Psychological review*, 1958, 65(6): 386.
- [4] Minsky M, Papert S. *Perceptrons* [M]. Cambridge, MA: MIT Press, 1969.
- [5] Bryson A, Denham W, Dreyfuss S. Optimal Programming Problem with Inequality Constraints, I: Necessary Conditions for Extremal Solutions [J]. *AIAA Journal*, 1963, 1: 25-44.

- [6] Linnainmaa S. The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors. Master's thesis, Univ. Helsinki, 1970.
- [7] Werbos P. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Boston: Harvard University, 1974.
- [8] Rumelhart D, Hinton G, Williams R. Learning Representations by Back-Propagating Errors[J]. Nature, 1986, 323: 533-536.
- [9] Hecht-Nielsen R. Theory of the Backpropagation Neural Network[C]. In International Joint Conferenceon Neural Networks(IJCNN), pages 593-605, 1989.
- [10] Hornik K, Stinchcombe M, White H. Multilayer Feedforward Networks are Universal Approximators[J]. Neural Networks, 1989, 2(5) : 359-366.
- [11] Hochreiter S. Untersuchungen zu Dynamischen Neuronalen Netzen. Diploma thesis, Institut fürInformatik, Technische Universität München, 1991.
- [12] LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W, Jackel L. Back- Propagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4) : 541-551.
- [13] Vapnik V. The Nature of Statistical Learning Theory[M]. Springer, New York, 1995. [Vapnik V. 统计学习理论的本质[M]. 张学工译. 北京: 清华大学出版社, 2000.]
- [14] Cortes C, Vapnik V. Support Vector Networks[J]. Machine Learning, 1995, 20: 273-297.
- [15] DeCoste D, Schoelkopf B. Training Invariant Support Vector Machines [J]. Machine Learning, 2002, 46: 161-190.
- [16] Friedman J. On Bias, Variance, 0/1—Loss, and the Curse- of- Dimensionality [J]. Data Mining and Knowledge Discovery, 1997, 1(1) : 55-77.
- [17] Viola P, Jones M. Robust Real-Time Face Detection[J]. International Journal of Computer Vision, 2004, 57(2) : 137-154.
- [18] Tenenbaum J, Silva V, Langford J. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science 290, 2000, 2319-2323.
- [19] Roweis S, Saul L. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science 290, 2000, 2323-2326.
- [20] Lowe D. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 2(60) : 91-110.
- [21] Hinton G, Salakhutdinov R. Reducing the Dimensionality of Data with Neural Networks[J]. Science 313, 2006, 504-507.
- [22] Smolensky P. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. chapter Information Processing in Dynamical Systems: Foundations of Harmony Theory, pages194-281. MIT Press, Cambridge, MA, USA, 1986.
- [23] Hinton G, Sejnowski T. Learning and Relearning in Boltzmann Machines. In Parallel Distributed Processing, volume 1, pages 282-317. MIT Press, 1986.
- [24] <http://www.kdnuggets.com/2014/02/exclusive-yann-lecun-deep-learning-facebook-ai-lab.html>.
- [25] 10 Breakthrough Technologies 2013. MIT Technology Review, 2013, 04, 23.
- [26] Bengio Y, Courville A, Vincent P. Representation Learning: A Review and New Perspectives[J]. IEEE Transactionson Pattern Analysisand Machine Intelligence, 2013, 35(8) : 1798-1828 , AUGUST 2013.
- [27] Lee H, Grosse R, Ranganath R, Ng A. Convolutional Deep Belief Networks for Scalableunsupervised

- Learning of Hierarchical Representations [C]. In Proceedings of the 26th International Conference on Machine Learning(ICML) , pages 609-616 , 2009.
- [28] Lee H. Deep Learning Methods for Vision. CVPR 2012 Tutorial.
- [29] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient- Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE , November 1998.
- [30] Hinton G, Osindero S, Teh Y. A Fast Learning Algorithm for Deep Belief nets[J]. Neural Computation , 2006 , 18(7) : 1527-1554.
- [31] Roux N, Bengio Y. Representational Power of Restricted Boltzmannmachines and Deep Belief Networks[J]. Neural Computation , 2008 , 20(6) : 1631-1649.
- [32] Lee H, Ekanadham C, Ng A. Sparse Deep Belief Net Model for Visual Area V2 [J]. In Advances in Neural Information Processing Systems(NIPS) , 2008 , 7 : 873-880.
- [33] Courville A, Bergstra J, Bengio Y. Unsupervised Models of Imagesby Spike- and- Slab RBMs [C]. In Proceedings of the Twenty-eight International Conference on Machine Learning(ICML'11) , 2011.
- [34] Salakhutdinov R. Deep Learning. CVPR 2012 Tutorial.
- [35] Salakhutdinov R, Hinton G. Deep Boltzmann Machines [C]. In Proceedingsof The Twelfth International Conference on Artificial Intelligence andStatistics(AISTATS'09) , vol. 5 , pages 448-455 , 2009.
- [36] Salakhutdinov R, Hinton G. A better Way to Pretrain Deep Boltzmann Machines [J]. In Advances in Neural Information Processing Systems , pages 2456-2464 , 2012.
- [37] Goodfellow I, Mirza M, Courville A, Bengio Y Multiprediction Deep Boltzmann Machines[J]. In Advances in Neural Information Processing Systems , pages 548-556 , 2013.
- [38] Ballard D. Modular Learning in Neural Networks[C]. In Proc. AAAI , pages 279-284 , 1987.
- [39] Ranzato M. Neural nets for vision. CVPR 2012 Tutorial.
- [40] Vincent P, Larochelle H, Bengio Y, Manzagol P. Extracting and Composing Robust Features with Denoising Autoencoders [C]. Proceedings of the Twenty- fifth International Conference on Machine Learning (ICML'08) , pages 1096-1103 , ACM , 2008.
- [41] Hinton G, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. arXiv preprint arXiv: 1207.0580 , 2012.
- [42] Krizhevsky A, Sutskever I, Hinton G. Imagenet Classification with Deep Convolutional Neural Networks [J]. In Advances in NeuralInformation Processing Systems , volume 1 , page 4 , 2012.
- [43] Ba J, Frey B. Adaptive Dropout for Training Deep Neural Networks[J]. In Advances in Neural Information Processing Systems , pages 3084-3092 , 2013.
- [44] Wan L, Zeiler M, Zhang S, LeCun Y, Fergus R. Regularization of Neural Network using Drop Connect , ICML 2013.
- [45] Wang S, Manning C. Fast Dropout Training[C]. In Proceedings of the 30th International Conferenceon Machine Learning(ICML-13) , pages 118-126 , 2013.
- [46] Baldi P, Sadowski P. Understanding Dropout[J]. In Advances in Neural Information Processing Systems , pages 2814-2822 , 2013.
- [47] Wager S, Wang S, Liang P. Dropout Training as Adaptive Regularization[C]. NIPS 2013.
- [48] Wager S, Fithian W, Wang S, Liang P. Altitude Training: Strong Bounds for Single- Layer Dropout. arXiv: 1407.3289v1.
- [49] Maaten L, Chen M, Tyree S, Weinberger K. Learning with Marginalized Corrupted Features [C].

ICML 2013.

- [50] Chen M, Weinberger K, Sha F, Bengio Y. Marginalized Denoising Auto-encoders for Nonlinear Representations [C]. ICML 2014.
- [51] Goodfellow I, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout Networks[C]. In International Conference on Machine Learning 2013.
- [52] Dean J, Corrado G, Monga R, Chen K, Devin M, Le Q, Mao M, Ranzato M, Senior A, Tucker P. Large Scale Distributed Deep Networks[J]. In Advances in Neural Information Processing Systems, pages 1232-1240, 2012.
- [53] Chellapilla K, Puri S, Simard P. High Performance Convolutional Neural Networks for Document Processing [C]. In International Workshop on Frontiers in Handwriting Recognition, 2006.
- [54] Raina R, Madhavan A, Ng A. Large-Scale Deep Unsupervised Learning using Graphics Processors[C]. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML), pages 873-880. ACM, 2009.
- [55] Ciresan D, Meier U, Masci J, Gambardella L, Schmidhuber J. Flexible, High Performance convolutional Neural Networks for Image Classification [C]. In Intl. Joint Conference on Artificial Intelligence IJCAI, pages 1237-1242, 2011.
- [56] Ngiam J, Coates A, Lahiri A, Prochnow B, Le Q, Ng A. On Optimization Methods for Deep Learning [C]. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 265-272, 2011.
- [57] Coates A, Huval B, Wang T, Wu D, Ng A. Deep Learning with COTS HPC Systems[C]. ICML 2013.
- [58] Arora S, Bhaskara A, Ge R, Ma T. Provablebounds for Learning Some Deep Representations[C]. In Proceedings of The 31st International Conference on Machine Learning, pages 584-592, 2014.
- [59] Sermanet P, LeCun Y. Traffic Sign Recognition with Multi-Scale Convolutional Networks[C]. Proceedings of International Joint Conference on Neural Networks, San Jose, California, USA, July 31- August 5, 2011.
- [60] Dahl G, Yu D, Deng L, Acero A. Context Dependent Pretrained Deep Neural Networks for Large Vocabulary Speech Recognition [J]. IEEE Trans. on Audio, Speech, and Language Processing, 2012, 20(1) : 30-42.
- [61] Le Q, Ranzato M, Monga R, Devin M, Corrado G, Chen K, Dean J, Ng A. Building High- Level Features Using Large Scale Unsupervised Learning[C]. In Proc. ICML 2012.
- [62] Ciresan D, Schmidhuber J. Multi-column Deep Neural Networks for Offline Handwritten Chinese Character Classification. Technical report[C]. IDSIA, 2013.
- [63] Graves A, Mohamed A, Hinton G. Speech Recognition with Deep Recurrent Neural Networks [C]. In Proc. ICASSP 2013.
- [64] Bluche T, Louradour J, Knibbe M, Moysset B, Benzeghiba F, Kermorvant C. The A2iA Arabic Handwritten Text Recognition System at the OpenHaRT2013 Evaluation. In International Workshop on Document Analysis Systems, 2014.
- [65] Ciresan D, Giusti A, Gambardella L, Schmidhuber J. Mitosis Detection in Breastcancer Histology Images with Deep Neural Networks. In Proc. MICCAI, volume 2, pages 411-418, 2013.
- [66] Girshick R, Donahue J, Darrell T, Malik J. Rich Feature Hierarchies for Accurate Objectdetection and Semantic Segmentation. Technical Report arxiv. org/abs/1311. 2524, UC Berkeley and ICSI, 2013.

- [67] Farabet C, Couprie C, Najman L, LeCun Y. Learning Hierarchical Features for Scenelabeling[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1915-1929.
- [68] Szegedy C, Toshev A, Erhan D. Deep Neural Networks for Object Detection[J]. In Advances in Neural Information Processing Systems, pages 2553-2561, 2013.
- [69] Khan S, Bennamoun M, Sohel F, Togneri R. Automatic Feature Learning for Robust Shadow Detection [C]. In IEEE Conference on Computer Vision and Pattern Recognition CVPR, 2014.
- [70] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li F. Large-scale Video Classification with Convolutional Neural Networks[C]. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.
- [71] Mnih V, Hinton G. Learning to Label Aerial Images from Noisy Data[C]. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), pages 567-574, 2012.
- [72] Lena P, Baldi P, Nagata K, Bartlett P, Pereira F., Burges C, Bottou L, Weinberger K. Deep Spatio-Temporal Architectures and Learning for Protein Structure Prediction. In Advances in Neural Information Processing Systems, pages 521-529, 2012.
- [73] Bengio Y, Ducharme R, Vincent P, Jauvin C. A Neural Probabilistic Language Model[J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
- [74] Mikolov T, Karafiat M, Burget L, Cernocký J, Khudanpur S. Recurrent Neural Network Based Language Model. in Proc. INTERSPEECH, 1045-1048, 2010.
- [75] Frinken V, Zamora-Martinez F, Espana-Boquera S, Castro-Bleda M, Fischer A, Bunke H. Long-Short Term Memory Neural Networks Language Modeling for Handwriting Recognition. In Pattern Recognition (ICPR), pages 701-704, 2012.
- [76] Yu K. Large-Scale Deep Learning at Baidu. CIKM 2013: 2211-2212.
- [77] Lin Y, Zhang T, Zhu S, Yu K. Deep coding network[C]. In Proceedings of the 27th International Conference on MachineLearning (ICML-10), pages 1405-1413, 2010.
- [78] Yu K, Lin Y, Lafferty J. Learning image representations from the pixel level via hierarchical sparse coding. CVPR 2011: 1713-1720.
- [79] Ji S, Xu W, Yang M, Yu K. 3d convolutional neural networks for human action recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2013, 35(1): 221-231.
- [80] Lu Z, Li H. A deep architecture for matching short texts. In Advances in Neural Information Processing Systems, pages 1367-1375, 2013.
- [81] Yu D, Seide F, Li G. Conversational speech transcription using context-dependent deep neural networks. In ICML, 2012.
- [82] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10, 000 classes[C]. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014.
- [83] Luo P, Tian Y, Wang X, Tang X. Switchable Deep Network for Pedestrian Detection[C]. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014.
- [84] Li W, Zhao R, Xiao T, Wang X. DeepReID: Deep Filter Pairing Neural Network for Person Re-Identification [C]. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014.
- [85] Ouyang W, Chu X, Wang X. Multi-source Deep Learning for Human Pose Estimation[C]. IEEE Conf. on Computer Vision and Pattern Recognition, June 2014.

- [86] Ouyang W, Wang X. Joint Deep Learning for Pedestrian Detection [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV) 2013.
- [87] Sun Y, Wang X, Tang X. Hybrid Deep Learning for Face Verification [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV) 2013.
- [88] Zhu Z, Luo P, Wang X, Tang X. Deep Learning Identity Preserving Face Space [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV) 2013.
- [89] Luo P, Wang X, Tang X. A Deep Sum-Product Architecture for Robust Facial Attributes Analysis [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV) 2013.
- [90] Luo P, Wang X, Tang X. Pedestrian Parsing via Deep Decompositional Neural Network [C]. In Proceedings of IEEE International Conference on Computer Vision (ICCV) 2013.
- [91] Sun Y, Wang X, Tang X. Deep Convolutional Network Cascade for Facial Point Detection [C]. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2013.
- [92] Ouyang W, Zeng X, Wang X. Modeling Mutual Visibility Relationship with a Deep Model in Pedestrian Detection [C]. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2013.
- [93] Luo P, Wang X, Tang X. Hierarchical Face Parsing via Deep Learning [C]. Prof. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2012.
- [94] Zhu Z, Luo P, Wang X, Tang X. Deep Learning Multi-View Representation for Face Recognition. arXiv: 1406.6947 2014.
- [95] Li S, Liu Z, Chan A. Heterogeneous Multi-task Learning for Human Pose Estimation with Deep Convolutional Neural Network, CVPR DeepVision Workshop 2014.
- [96] Jin J, Fu K, Zhang C. Traffic Sign Recognition With Hinge Loss Trained Convolutional Neural Networks. Intelligent Transportation Systems, 2014.
- [97] 胡振, 傅昆, 张长水. 基于深度学习的作曲家分类问题, [J]. 计算机研究与发展, 2014.
- [98] Li P, Liu Y, Sun M, Izuha T, Zhang D. A Neural Reordering Model for Phrase-based Translation. In Proceedings of COLING 2014, Dublin, Ireland, August.
- [99] Li P, Liu Y, Sun M. Recursive Autoencoders for ITG-based Translation. In Proceedings of EMNLP 2013, Seattle, Washington, USA, October.
- [100] Hu X, Zhang J, Li J, Zhang B. Sparsity-regularized HMAX for visual recognition [J]. PLOS ONE, 2014, 9(1): 1-12.
- [101] Shi T, Liang M, Hu X. A reverse hierarchy model for predicting eye fixations [C]. Proc. of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, USA, June 24-27, 2014.
- [102] Chen N, Zhu J, Chen J, Zhang B. Dropout Training for Support Vector Machines. AAAI 2014.
- [103] Xie J, Xu L, Chen E. Image denoising and inpainting with deep neural networks. In Advances in Neural Information Processing Systems, pages 350-358, 2012.
- [104] Li X, Wu X. Labeling unsegmented sequence data with DNN-HMM and its application for speech recognition [C]. In Proc. ISCSLP 2014.
- [105] Li X, Wu X. Decision tree based state tying for speech recognition using DNN derived embeddings. In Proc. ISCSLP 2014.
- [106] Gong C, Li X, Wu X. Recurrent neural network language model with part-of-speech for mandarin speech

- recognition. in Proc. ISCSLP 2014.
- [107] Huang W, Song G, Hong H, Xie K. Deep Architecture for Traffic Flow Prediction: Deep Belief Nets with Multi-task Learning. IEEE Transaction on Intelligent Transportation Systems, 2014.
- [108] Bai Y, Yang K, Yu W, Ma W, Zhao T. Learning High-level Image Representation for Image Retrieval via Multi-Task DNN using Clickthrough Data[C]. CoRR 2013.
- [109] Cui Q, Gao B, Bian J, Qiu S, Liu T. Learning Effective Word Embedding using Morphological Word Similarity. arXiv: 1407.1687, 2014.
- [110] Chen X, Xiang S, Liu C, Pan C. Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks[J]. IEEE Geoscience and Remote Sensing Letters(GRSL), 2014, 11(10): 1797-1801.
- [111] 陈雪云. 基于深层神经网络的遥感图像目标检测[C]. 博士学位论文, 中科院自动化研究所, 2014, 北京.
- [112] Huang Y, Wang W, Wang L, Tan T. Multi- task deep neural network for multi- label learning. IEEE International Conference on Image Processing(ICIP), pp. 2897-2900, Melbourne, Australia, September 15-18, 2013.
- [113] Yi D, Lei Z, Liao S, Li Z. Deep Metric Learning for Person Re-Identification[C]. In Proceedings of International Conference on Pattern Recognition(ICPR), Sweden, August 24-28, 2014.
- [114] Wang W, Cui Z, Chang H, Shan S, Chen X. Deeply Coupled Auto-encoder Networks for Cross-view Classification. arXiv: 1402.2031, 2014.
- [115] Si Y, Zhang Q, Li T, Pan J, Yan Y. Prefix tree based n-best list re-scoring for recurrent neural network language model used in speech recognition system. in Proc. INTERSPEECH 2013.
- [116] Gao W, Zhou Z. Dropout Rademacher Complexity of Deep Neural Networks. arXiv: 1402.3811, 2014.
- [117] 余凯, 贾磊, 陈雨强, 徐伟. 深度学习的昨天、今天和明天[J]. 计算机研究与发展, 2013, 50(9): 1799-1804.
- [118] Anselmi F, Leibo J, Rosaco L, Mutch J, Tacchetti A, Poggio T. Unsupervised Learning of invariant representations in hierarchical architectures. CBCL paper, MIT, Sept. 27, 2013.
- [119] Ladicky L, Russell C, Kohli P, Torr P. Associative hierarchical crfs for objectclass image segmentation. In: ICCV, 2009.
- [120] George D, Hawkins J. Towards a mathematical theory of cortical micro-circuits[J]. PloS Computational Biology, Vol. 5, No. 10, 2009
- [121] Síma J. Loading deep networks is hard[J]. Neural Computation, 1994, 6(5): 842-850.
- [122] Síma J. Training a single sigmoidal neuron is hard. Neural Computation, 2002, 14(11): 2709-2728.
- [123] Windisch D. Loading deep networks is hard: The pyramidal case[J]. Neural Computation, 2005, 17(2): 487-502.
- [124] 张长水. 机器学习面临的挑战[J]. 中国科学: 信息科学, 2013 43(12): 1612-1623.
- [125] Bengio Y. Practical recommendations for gradient-based training of deep architectures. In K. - R. Müller, G. Montavon, and G. B. Orr, editors, NeuralNetworks: Tricks of the Trade. Springer, 2013.
- [126] Kurzweil R. How to Create a Mind: The Secret of Human Thought Revealed[M]. Penguin Books. 2013. (雷·库兹韦尔. 如何创造思维: 人类思想所揭示出的奥秘[M]. 盛杨燕, 译. 杭州: 浙江人民出版社, 2014.)
- [127] Bengio Y. Learning Deep Architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-127.

作者简介

封举富 生于1967年10月。北京大学信息科学技术学院智能科学系教授、博士生导师。1997年于北京大学数学学院获博士学位。主要研究领域为图像处理、模式识别与机器学习、生物特征识别等。在国内外期刊和重要国际会议上发表论文100多篇，包括IEEE Trans. PAMI、IEEE Trans. CSVT、IEEE Trans. SMC-B、Pattern Recognition、Neural Computation、ICML、COLT、CVPR等。2005年获教育部新世纪优秀人才支持计划。1993年，获第一届亚洲计算机视觉国际会议优秀论文奖。2000年获中国高校科技进步二等奖。2012年获公安部科学技术二等奖。现任中国计算机学会模式识别与人工智能专委会副主任委员。



王立威 生于1975年5月。北京大学信息科学技术学院智能科学系教授、博士生导师。2005年于北京大学数学学院获博士学位。主要研究兴趣为机器学习理论。在机器学习顶级会议NIPS、COLT、ICML和顶级期刊JMLR、IEEE Trans. PAMI发表论文多篇。其中2008年发表于机器学习理论最高会议COLT的论文On the Margin Explanation of Boosting Algorithms是中国大陆学者在该会议上的首篇论文。2010年入选AI's 10 to Watch，是首位获得该奖项的亚洲学者。2012年获得国家自然科学基金优秀青年基金。2012年获教育部新世纪优秀人才支持计划。现任Journal of Computer Science and Technology (JCST)等期刊编委。中国计算机学会模式识别与人工智能专委会委员。



胡占义 1961年生于山西。中国科学院自动化研究所研究员、博士生导师。1993年于比利时列日大学获计算机视觉专业国家博士学位。主要研究领域为计算机视觉，在计算机视觉领域重要国际期刊和国际会议上发表论文150多篇。现为《中国科学》、《科学通报》、《Journal of Computer Science and Technology》编委，《计算机辅助设计与图形学学报》副主编。先后担任ICCV Local Co-Chair、ACCV 2012 Program Co-Chair。2004年获国家自然科学二等奖。



基于搜索的软件工程研究进展与趋势

CCF 软件工程专业委员会

李 征¹ 巩敦卫² 聂长海³ 江 贺⁴

¹北京化工大学信息科学与技术学院，北京

²中国矿业大学信息与电气工程学院，徐州

³南京大学计算机科学与技术系，南京

⁴大连理工大学软件学院，大连

摘要

软件工程自从 1968 年提出以来，一直是提高软件开发效率，保障软件质量的有效手段。基于搜索的软件工程（Search Based Software Engineering, SBSE）是传统软件工程和智能计算（Intelligent Computing）交叉的新兴研究领域，它采用智能计算领域的现代启发式搜索优化算法解决软件工程相关问题，核心是实现智能化和自动化的软件工程相关问题求解，被 2007 年度 IEEE 国际软件工程大会（ICSE）确立为软件工程的未来发展方向之一。基于搜索的软件工程已经在软件测试数据自动生成、程序错误自动修复等方面取得显著的研究成果，有效地促进了软件工程学科的发展。报告将从两个方面论述国内外发展现状：一方面是智能计算领域相关技术综述，特别是已经应用到软件工程领域的相关演化算法的综述；另一方面，从软件工程的各个阶段（包括需求分析、设计、测试、维护等）综述基于搜索的软件工程发展现状。

关键词：软件工程，智能计算

Abstract

Software Engineering was introduced in 1968, and it has been an effective way to improve the software development process and assure the software quality. Search based software engineering (SBSE) is a promising combination research area of the traditional software engineering and intelligent computing, which using meta-heuristic search algorithms to solve different optimization problems in diverse software engineering areas. SBSE approaches can be implemented automatically and intelligently to obtain solutions for complex tasks and it has been recognized as a promising future of the software engineering. There have been many exciting research achievements in SBSE, such as automatic test case generation, automatic bugs fixing. This report will introduce the SBSE in two aspects, one is the survey of the search algorithms in intelligent computation, which have been applied in SBSE; the other is the art of state of the SBSE, including requirement, software design, testing and maintenance.

Keywords: software engineering, intelligent computation

1 引言

软件危机的出现，促进了软件工程学的形成和发展，并不断进化。从软件工程开发方法来看，人们先是从面向 01 代码的原始数字信息到面向过程（Procedure Oriented），再从面向过程提升到面向对象（Object Oriented），再从面向对象进化到面向服务（Service Oriented）、面向方面（Aspect Oriented）、面向领域（Domain Oriented）等。软件工程的发展历史本身就是一个不断提出问题又不断解决问题的过程。

基于搜索的软件工程（Search-Based Software Engineering, SBSE^①）是将传统的软件工程问题转化为基于搜索的优化问题，并使用现代启发式搜索算法^②（meta-heuristic search algorithms，也称元启发式算法）解决问题的研究和实践方法。相比启发式搜索算法，现代启发式搜索算法定义智能搜索策略，增强启发式算法搜索性能，以一种智能形式在问题的解空间中搜索最优解或近似最优解。

SBSE 最早可以追溯到 1976 年，Webb Miller 和 David Spooner 尝试把优化算法用于浮点测试数据的生成^[1]。1992 年，Xanthakis 和他的同事首次将搜索算法用于解决软件工程问题^[2]。2001 年，Harman 和 Jones 正式提出将软件工程问题转化为基于搜索的优化问题，并采用以遗传算法（Genetic Algorithm）、模拟退火算法（Simulated Annealing Algorithm）、禁忌搜索算法（Tabu Search Algorithm）等为代表的现代启发式搜索算法来求解^[3]，这可以说是奠定了基于搜索的软件工程的里程碑。面对规模日益庞大和复杂的软件，传统的软件工程方法已经不能有效解决软件开发过程中的问题。基于搜索的软件工程的提出，为解决这些问题提供了新的解决思路，因此被 2007 年的 IEEE 国际软件工程大会正式确立为软件工程领域未来发展的新方向^[4]。

传统的软件工程的解决问题方法是在问题空间通过算法来构造一个解，而基于搜索的软件工程是在解空间（所有可能的解）中使用启发式搜索算法以具体问题的适应值函数作为向导搜索最优解。通常，使用基于搜索的优化算法解决问题，需要满足以下两个条件^[3,4]：

1) 设计出问题解决方案表达方式（Solution Representation）：对所需解决问题的结果，必须能通过相应的编码表示出来，以构成搜索算法中的染色体，进行相应的运算。

2) 设计出相应的适应度函数（Fitness Function）：对解进行评价，比较不同解之间的优劣。在搜索解空间内，适应度函数可以指引搜索的方向，寻找满足条件的区域。

Harman 在文献[4]和[5]中曾表述，软件工程师对他们的问题已经有一个合适的呈现，并且许多软件工程师在软件度量方面所取得的成果，也为构造适应度函数打下了基础。

① 后续内容将“基于搜索的软件工程”简称为“SBSE”。

② 后续内容中提到的“搜索算法”或“搜索技术”通常指“现代启发式搜索算法或技术”。

由于基于搜索的软件工程解决问题的方法主要由以上两步组成，针对任何问题，只要能够设计出问题解的表示方式和适应度函数，就可以应用该方法，因此具有很强的普适性，可以方便地应用到不同领域问题上。另外，针对同一问题的不同规模，基于搜索的方法求解方式是不变的，因此容易扩展到大规模的工程问题求解上去。基于搜索的软件工程方法避免了传统方法的一些不足，能在尽可能降低成本的前提下尽可能高效地自动化完成软件开发任务，是软件工程领域中一种高效实用的新方法。

经过十几年的发展，基于搜索的软件工程研究已经取得了显著的成绩。该领域的创始人 Harman 教授所在的研究团队创建并维护了一个 SBSE 研究领域的文献库，收录主要国际期刊和国际会议关于 SBSE 的发表文献。根据该文献库统计数据，截至 2013 年，从事 SBSE 研究团体已超过 270 个，遍布全球 40 多个国家，超过 800 名研究人员，直接相关的发表文章超过 1000 篇[⊖]。

图 1 显示了 1976 ~ 2011 年期间基于搜索的软件工程领域发表文章情况，可以看出近 10 年发表文章数量呈现显著增长趋势。

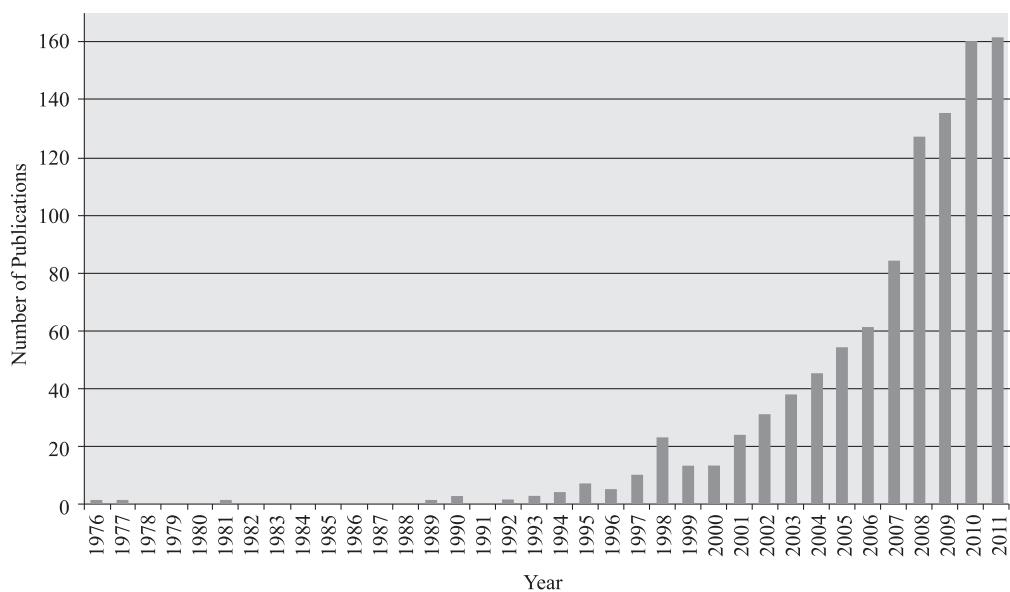


图 1 SBSE 领域发表文献趋势

基于搜索的软件工程已经形成一个独立的研究方向，软件工程领域主流学术期刊，包括顶级 IEEE Transactions on Software Engineering (TSE) 在内，都相继出版了该领域的专刊，具体如表 1 所列。

⊖ The repository of publications in SBSE (<http://www.sebase.orgsbsepublications/>) .

表 1 基于搜索的软件工程国际期刊专刊

年度	期刊名称
2007	Journal of Software Maintenance and Evolution (JSME)
2008	Computers and Operations Research (COR)
2010	IEEE Transactions on Software Engineering (TSE)
2010	The journals Information and Software Technology (IST)
2011	Software: Practice and Experience (SPE)
2011	EMpirical Software Engineering (EMSE)
2013	Journal of Systems and Software (JSS)

从研究内容上看，基于搜索的软件工程已经应用到软件开发生命周期的各个阶段，从测试需求和工程规划，到软件的维护和重建。本文将在后续的章节中详细介绍基于搜索的软件工程在生命周期不同阶段的具体应用。

2 国际研究现状

SBSE 是智能计算与软件工程的交叉，报告将从两个方面论述国内外发展现状：一方面是智能计算领域的相关技术综述，特别是已经应用到软件工程领域相关智能优化方法的综述；另一方面重点从软件工程的各个阶段（包括需求分析、设计、测试、维护等）综述 SBSE 的发展现状并结合具体应用（包括应用场景、应用效果等）论述 SBSE 在工业界的应用现状。

2.1 智能优化方法

图 2 统计了在 SBSE 领域发表文献中涉及的各种智能优化算法使用的频率，可以看出遗传算法（GA）是最广泛使用的优化算法，被 300 多篇 SBSE 领域文献的相关研究使用。

本节主要介绍一些广泛使用的智能优化算法，包括遗传算法、爬山算法、模拟退火算法、蚁群算法和粒子群算法等。

2.1.1 遗传算法

在现有的智能优化算法中，遗传算法（Genetic Algorithm, GA）是软件工程领域中使用最为广泛的一种算法^[6]。

遗传算法是一种模拟自然界生物进化过程的启发式搜索算法。这种启发式算法能够产生一个针对问题解的群体并对其进行进化和优化。遗传算法属于演化算法的一种，这类算法通过模拟自然选择的过程对产生的解进行优化。优化操作包括选择、变异和交叉。

在遗传算法中，最优化问题产生的一系列候选解组成了种群，种群的不断演化能够得到更好的解。每个候选解都有不同的特征，演化过程是通过候选解特征间的变异和交叉实现的。

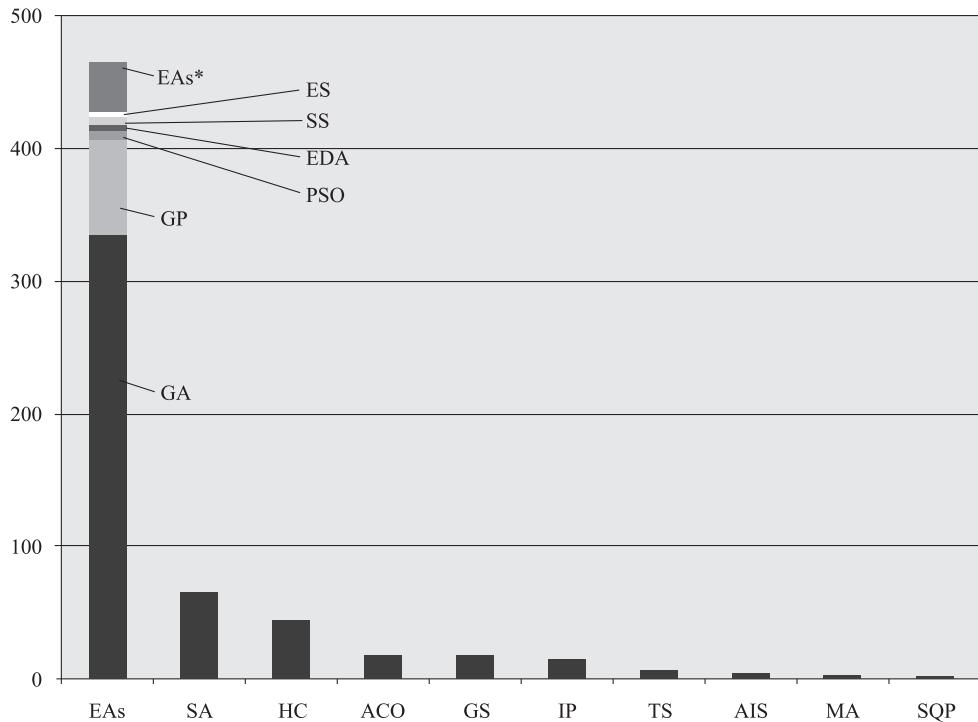


图 2 SBSE 研究领域使用的主要搜索算法

演化过程通常由一组随机产生的个体构成初始种群，然后对种群不断迭代，每一次迭代过程中的种群被称作一代群体。在每一代中，对群体的每个个体进行评价，度量函数为解的个体的适应度函数。经过选择、交叉、变异等操作，具有较好评估值的个体会被选择至下一代种群中，这样每一代不断优化产生下一代种群。迭代的终止条件是达到迭代次数或者是达到评估函数的目标值。

遗传算法在整个演化过程中有许多变化，关键在于目标函数引导搜索过程的方式、重组和基于种群的搜索过程。另一种独立于遗传算法的演化计算形式为演化策略 (evolution strategy)，研究者证明其在测试数据生成方面优于遗传算法^[7]。其他遗传算法的变种包括遗传编程 (genetic programming)、粒子群优化算法 (particle swarm optimization)、进化规划 (evolutionary programming)、进化策略 (evolutionary solution) 等。各种演化算法已成功应用于基于搜索的软件工程中，包括制定能够捕捉软件项目的预测模型^[8,9]和软件测试中的应用^[10]。

2.1.2 爬山算法

爬山算法 (Hill Climbing, HC) 是一种局部搜索算法。它从问题的某个可行解出发，

通过每次更改解的一个决策变量以寻找更好的解。根据更改决策变量方式的不同，爬山算法可以分为两种，即近邻爬山算法（next ascent hill climbing）和最陡爬山算法（steepest ascent hill climbing）。在近邻爬山算法中，更改首位决策变量以寻找更优解；在最陡爬山算法中，搜索所有决策变量的临近解以寻找更优解。如果对决策变量的更改能够得到一个更好的解，那么就以更改的决策变量进行再次搜索，重复此过程直至解质量无法再提高。

爬山算法从搜索空间中的一个解出发，通过不断迭代，最终可达到一个局部最优解。算法停止时得到的解的质量依赖于算法的初始解的选取、邻域选点的规则和算法的终止条件等。爬山算法作为一个简单高效的搜索算法已广泛应用于基于搜索的软件工程领域中^[11,12]。

2.1.3 模拟退火算法

模拟退火（Simulated Annealing, SA），也叫做蒙特卡罗退火，是源于对热力学中退火过程的模拟。在给定某一温度下，通过缓慢下降温度参数，从而增加退火强度。模拟退火算法可以视为爬山算法的一个变种，通过允许当前解移动到非最优个体来解决爬山算法容易陷入局部最优问题。这种算法首先选取搜索空间中的一个可行解作为搜索起始点，迭代过程中每一步先选择一个邻域点，然后计算从现有位置到达邻域居的概率。

模拟退火算法新解的产生和接受可分为以下步骤：

- 1) 当前解经过简单地变换产生新的解，其中变换包括对构成新解的全部或部分元素进行置换、互换等。
- 2) 计算当前解的目标函数值与新解所对应的目标函数值的差。
- 3) 判断新解是否被接受，判断的依据是一个接受准则，例如最常用的接受准则是Metropolis 准则。
- 4) 当新解被确定接受时，就用新解代替当前解同时修正目标函数值。此时，实现了对当前解的一次迭代。

模拟退火算法具有渐近收敛性，在理论上已被证明它是一种以概率 1 收敛于全局最优解的全局优化算法。该算法已被广泛应用于基于搜索的软件工程的领域中^[12,13,14,15]。

2.1.4 蚁群算法

蚁群算法（Ant Colony Optimization, ACO），又称蚂蚁算法，是一种用来在图中寻找优化路径的机率型算法。

蚁群算法的提出借鉴和吸收了现实世界蚂蚁集体寻径的行为特征。蚂蚁觅食过程中分泌一种信息素的物质，该物质随时间不断挥发。蚂蚁利用信息素作为媒介进行信息沟通，一条路径上留下的信息素浓度的大小与这条路径上通过的蚂蚁数成正比。当通过的蚂蚁越多，留下的信息素越多，导致后来蚂蚁选择该条路径的概率提高，从而建立最短的移动路径。这些规则综合起来具有两个方面的特点：多样性和正反馈。其

中多样性保证了蚂蚁在觅食的过程不会走进死胡同而无限循环；正反馈机制则保证了相对优良的信息能够保存下来。蚁群算法已成功应用于基于搜索软件工程领域的软件测试中^[16~18]。

2.1.5 粒子群算法

粒子群算法（Particle Swarm Optimization，PSO）是一种进化计算技术，是通过模拟鸟群觅食过程中的迁徙和群聚行为而提出的一种基于群体智能的全局随机搜索算法。粒子群算法将群体中的个体看做是在搜索空间中没有质量和体积的粒子，每个粒子以一定的速度在解空间运动，并向自身历史最佳位置和邻域历史最佳位置聚集，实现对候选解的优化。

粒子群算法随机选择一群粒子作为初始种群，然后通过迭代找到最优解。所有的粒子都有一个适应值，每个粒子具有运动方向和距离。在每一次迭代过程中，粒子通过跟踪两个极值来更新自身：第一个极值是粒子本身所找到的最优解，这个极值称为个体极值；第二个极值是整个种群目前找到的最优解，这个极值称为全局极值。迭代此过程直至达到全局最优解。

粒子群算法通过粒子间的竞争和协作以实现在复杂搜索空间中寻找全局最优解的目的，它具有易理解、易实现、全局搜索能力强等特点，已广泛应用于基于搜索的软件工程领域中^[19~21]。

2.1.6 遗传编程

遗传编程（Genetic Programming，GP）又称基因编程，是一种进化计算技术，它能够从较高层次状态自动解决问题，自动地进行问题求解，不需要预知或指定解的形式及结构，是一种从生物演化过程得到灵感的自动化生成和选择计算机程序来完成用户定义的任务的技术。从理论上讲，用户使用遗传编程只需要告诉计算机“需要完成什么”，而不用告诉它“如何去完成”，最终可能实现真正意义上的人工智能。

遗传编程在1992年由美国John Koza正式提出^[22]，文中用层次化的结构性语言描述问题，利用树形结构来表达计算机程序，其中包括函数集F（节点）和终止符集T（叶子）。函数集F包含若干个函数，终止符集T包含若干变量或常量，利用基本的遗传操作（如选择和交叉）产生新个体，并采用个体适应度来评价个体的优劣。

由于遗传编程所需的计算量非常之大（处理大量候选的计算机程序），因此在20世纪90年代，人们只能用它来解决一些简单的问题。2000年以来，随着遗传编程技术自身的发展和中央处理器计算能力的指数级提升，GP开始产生了一大批显著的成果，在多个领域（如量子计算、电子设计、游戏比赛、排序、搜索等）均取得了重大发展，且使用范围也在不断扩展^[23~25]。

2.2 基于搜索的软件工程

图3显示了基于搜索的软件工程技术在软件工程生命周期各个阶段发表文章数量的

分布,可以看出超过50%的文章是基于搜索的软件测试与调错方向。本章将重点介绍基于搜索的软件测试、需求分析、软件设计、软件重构与维护和软件项目开发管理。

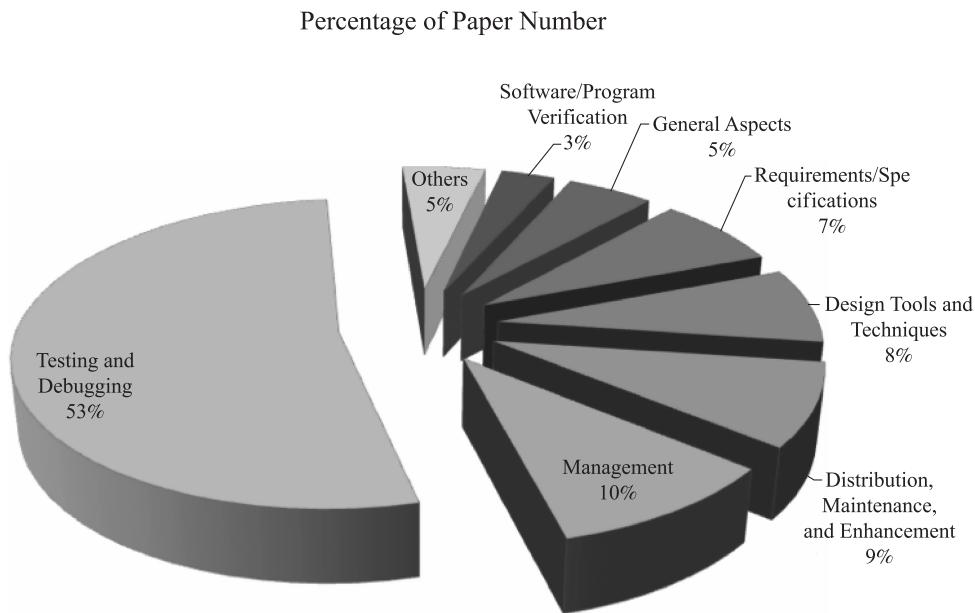


图3 SBSE在软件工程中研究领域的分布

2.2.1 软件测试

软件测试是软件质量的保证,为了在软件发布之前及时发现和修正软件中可能存在的缺陷,人们提出了很多软件测试方法。例如,结构测试可以检测程序中的各个语句、条件或分支等引发的错误;基于性质的软件测试技术可以有效检测软件系统是否被正确实现;基于组合覆盖的软件测试可以有效检测软件系统中各种因素相互作用引发的故障;变异测试不仅能够检测软件可能潜在的各种错误,还可以对已有的测试质量进行评估,并为进一步测试提供依据。

软件测试的目的是度量和提高软件质量,通过对待测软件(Software Under Test, SUT)及其相关文档、测试标准进行分析,进而设计并执行一系列的测试用例,测试人员往往期望能检测出软件中尽可能多的故障。在理想情况下,为了对测试的有效性和测试后的软件质量有一个较高的信心,软件测试应该是尽可能穷尽整个待测输入空间的,但在实际工程中并没有足够多的资源来执行穷尽测试所需的测试用例。因此,传统意义上的软件测试通常强调测试用例的精心设计,以期达到用最小的代价、最科学的方法,实现对待测试软件最为系统有效的测试。然而,在实践中,人们发现那些精心设计的测试用例不仅不容易自动化,而且有时恰恰是精心地绕过了软件中存在的故障。另一方面,高度自动化的随机测试在很多测试场景下表现得并不逊色,甚至有时比那些理论上很有效的测试技术还更加有效,此外随机测试在可靠性度量和统计估计上也具有天然的优势。但是,纯粹的随机测试却难以实现更深层次的测试,例如要实现语句覆盖、分支覆盖、

路径覆盖等各种覆盖充分性准则测试。

基于搜索的软件测试是基于搜索的软件工程领域的一个重要分支，是一种利用搜索技术来解决软件测试中各项问题的测试方法。从基于搜索的软件工程的角度来看，软件测试的核心过程就是利用各种方法搜索软件中存在的潜在错误的过程，这为搜索技术的应用提供了一个非常好的舞台。目前，基于搜索的软件工程中超过 50% 的研究关注的正是软件测试这一领域，尤其是如何利用搜索技术高效且自动化地为待测系统生成测试数据。McMinn^[26] 和 Ali 等人^[27] 曾分别对基于搜索的测试数据生成技术和实证研究进行了总结，Afzal 等人^[28] 则总结了搜索技术在软件非功能性测试上的应用。最近，McMinn^[29] 进一步讨论了基于搜索的软件测试的研究进展和发展趋势。

基于搜索的软件测试最早可以追溯到 1976 年，美国学者 Miller 和 Spooner 提出了一种利用遗传算法生成浮点测试数据的测试方法^[1]。与当时广泛使用的基于符号执行或约束求解的测试数据生成方法不同，这个新方法使用一个评价函数来评估当前测试数据的执行是否接近于测试人员所期望的预期路径，从而使得测试数据的生成问题也就转化为评价函数的优化问题。在遗传算法的优化作用下，那些与预期路径“距离”较远的测试数据将被丢弃，而更加符合预期路径的测试数据将被最终用于测试执行。

Miller 和 Spooner 并没有在这一领域延续他们的工作，搜索技术也直到 1990 年才由 Korel 应用于测试数据的生成^[31]。随后的 1992 年，Xanthakis 进一步使用遗传算法来生成测试数据^[2]。在此之后，这一研究领域产生了爆炸式的发展，搜索技术被频繁且大范围地应用到软件测试中，成功地解决了诸如功能测试^[32,33]、执行时间测试^[34~40]、集成测试^[41~47]、压力测试^[48~53]、变异测试^[54~62]、组合测试^[63~75]、回归测试^[76~83] 等各种测试方法中的各种问题。

在各种测试方法中，测试数据生成是搜索技术在软件测试领域应用最为广泛的场景。例如，在结构测试中，当使用搜索技术生成覆盖特定分支或条件的测试数据时，待测程序将首先被插桩，并执行一些随机产生的测试数据。随后，这些测试数据的质量将依据适应值函数进行评估，在搜索机制的作用下，更加接近测试目标的测试数据将被不断生成，最终达到预期的测试目的。在这一过程中，适应值函数的设计是应用搜索技术求解软件测试问题的关键。分支距离和层接近度，以及两者之间的相互结合^[84] 是最为常用的几种方法。

分支距离度量最初由 Korel 使用，其通过改变变量的方法来生成测试数据。在执行测试数据时，如果一条非指定路径被执行，那么就会产生与指定路径间的偏差，这一偏差也称为分支距离，而搜索的目标，就是尽可能地减少这一偏差。例如，我们考虑如图 4 的一个三角形程序^[26]，并令预期的测试路径为 $\langle s, 1, 5, 9, 10, 11, 12, 13, 14, e \rangle$ 。如果程序执行了输入 ($a = 10, b = 20, c = 30$)，在节点 1 和 5 处将通过相应的预期分支，但在节点 9 处却不能按预期通过正确分支。

为了度量分支距离，我们假定分支的谓词均是 “ $a \text{ op } b$ ” 的形式，这里 a 和 b 是算术表达式， op 是关系运算符，而相应的目标函数为 “ $f \neq 0$ ” 的形式，表 2 给出了谓词表达式与目标函数之间的关系。当谓词取正确分支时，目标函数取负值，反之则取正值。如

果我们想要执行谓词的正确分支，就需要尽可能降低目标函数的值，对应地，测试数据生成问题也就转化为目标函数的最小化问题。例如，在节点 9 处，针对输入 ($a = 10$, $b = 20$, $c = 30$)，有 $f = c - b = 10$ 。如果我们想要执行该节点处的正确分支，就需要尽可能地设法降低 f 的值。

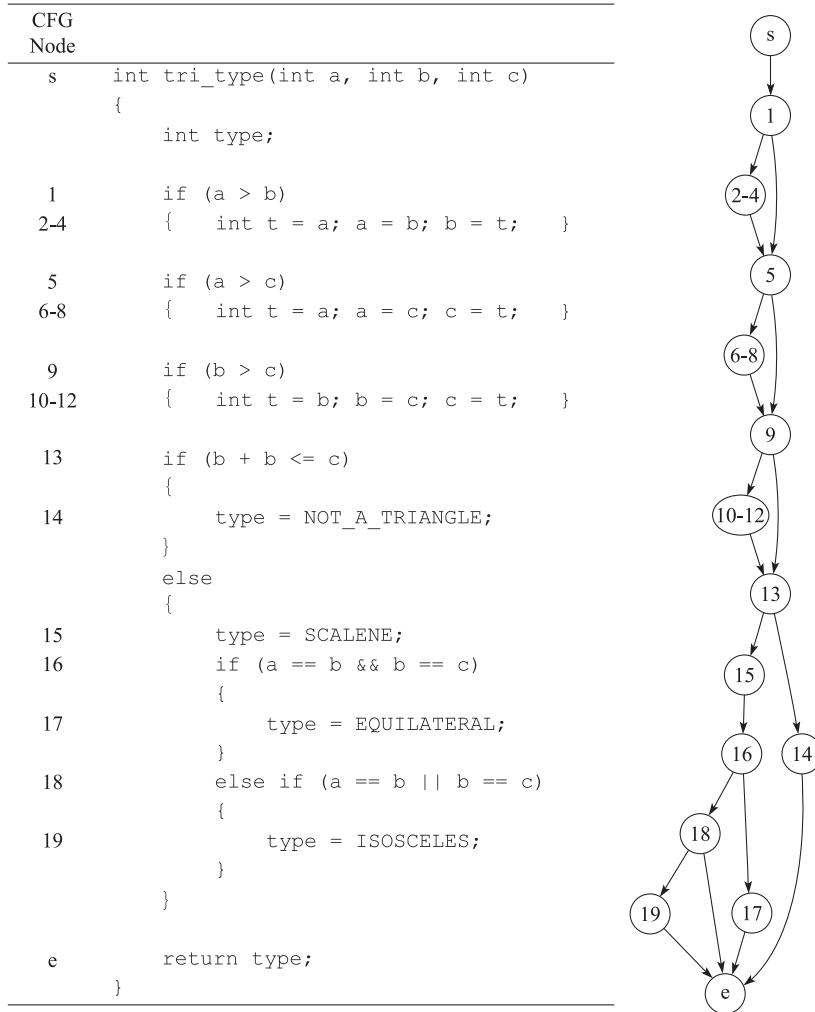


图 4 一个三角形示例程序

表 2 不同谓词表达式与目标函数之间的关系

谓词	目标函数	关系操作符
$a > b$	$b - a$	$<$
$a \geq b$	$b - a$	\leq
$a < b$	$a - b$	$<$
$a \leq b$	$a - b$	\leq
$a = b$	$abs(a - b)$	$=$
$a \neq b$	$-abs(a - b)$	$<$

为了在节点 9 处执行正确分支，可以通过改变变量的方法进行局部搜索，这一搜索过程旨在依次改变每个变量的值，同时保持其他变量值不变。改变变量值的第一个阶段称为探索阶段，通过增加或减少变量的初始值来对变量的邻域进行探索。如果在某个方向上的改变可以降低目标函数值，就可以以此得到某个模式。随后，在下一步模式阶段，就可以根据这个模式进行进一步的相似变换，直到找到目标函数的最小值为止。每个变量都将依据上述过程逐一进行优化，以搜索得到新的测试数据。

在上述例子中，程序执行过程中在节点 9 偏离预期路径，此时改变 a 的值并不会影响目标函数，所以首先选择变量 b 。可以发现，减少 b 的值会使目标函数更差，所以需要增加 b 的值，直到 $b > c$ 。例如，当 $b = 31$ 时新的输入 ($a = 10, b = 31, c = 30$) 就可以实现在节点 9 处按照预期路径执行。但是，这一输入在节点 13 处仍有偏离，此时需再次调用局部搜索。此时，为了在调整输入变量时维持前面的执行路径，新的目标函数将变为 $(a + b) - c$ 。由于减少 b 的值会违反已有的路径约束，而增加 b 的值可以改善目标函数，因此我们最后可以得到测试数据 ($a = 10, b = 40, c = 30$) 以执行预期路径。

然而，上述局部搜索的效果依赖于初始搜索的结果。例如，在图 5 所示的例子中^[26]，如果初始输入选择的是 ($a = 10, b = 10, c = 10$)，控制流将直接进入最后一个节点。此时若想要执行到目标语句，变量 c 的值就需要小于 0，但这样的话就会违反前面路径的执行条件。在这种情况下，局部搜索就失败了。在搜索技术的应用中，如果进入到类似的一个无法提高目标函数的变量值域，就会造成很多无效的搜索和浪费。为了提高搜索效率，可以利用一些来自程序的附加信息，特别是那些影响当前分支节点的变量信息。例如在图 4 中的节点 5 处，由于改变变量 a 和 b 的值可能会改变已成功通过节点 1 处的路径条件，因此搜索变量 c 比搜索变量 a 和 b 更好。

```
void nested_example(int a, int b, int c)
{
    if (a == b)
        if (b == c)
            if (c < 0)
                // target
}
```

图 5 局部搜索失效的例子

除了分支度量外，另一种计算适应值的方法是层接近度。假设当前的执行路径为 t ，预期的执行路径为 t^* ，那么就可以定义层接近度为 $a(t)/|t^*|$ ，其中 $a(t)$ 代表路径 t 没有经过 t^* 中的节点的个数， $|t^*|$ 代表预期路径的节点总数。例如，对于预期路径 $<s, 1, 5, 9, 10, 11, 12, 13, 14, e>$ 来说，由于输入 ($a = 10, b = 20, c = 30$) 将执行路径 $<s, 1, 5, 9, 13, 14, e>$ ，因此有层接近度为 $2/10 = 0.2$ 。显然，层接近度越小代表执行路径越接近预期情况，而测试数据的生成问题，也同样转化为目标函数的最小值问题。

适应值函数的设计是使用搜索技术生成测试数据的关键，不合适的适应值函数将不能有效地指导搜索的进行。例如，在上述测试数据生成中，对于仅包含一个布尔型变量

(也称为 flag 变量) 的分支语句, 其真假两个分支将使搜索空间形成两个高原, 其中一个非常符合搜索目标, 而另一个则非常不符合搜索目标。在这样的情况下, 由于缺少相应的梯度信息来指导搜索的进行, 搜索技术的应用将变得随机和低效^[85,86]。为了解决这个问题, Harman 等人^[87~89]提出了一种可测试性的转化方法。该方法将通过替换分支条件来为待测软件产生一个适合进行搜索的临时版本, 并以此来生成合适的测试数据。对于其他类似的情况, 例如基于状态的程序或者字节代码程序等, 相应的转化方法^[90~92]也被相继提出以提高软件的可测试性。

在测试数据生成中, 除了覆盖程序的某些特定结构外, 搜索技术同样可以依据其他不同的测试标准来为不同的测试方法生成测试数据。例如, 在功能测试中, 搜索技术的一个非常成功的应用是测试 DaimlerChrysler 的停车控制系统的一个早期版本^[32,33]。软件功能测试的目标是检测系统的逻辑行为是否与预期的规格说明相一致, 对于停车控制系统来说, 它将自动地沿侧方向将汽车移入车位, 并在这个过程中依赖不同的传感器来追踪汽车的运动轨迹, 从而在汽车与其他实体间距离过近时做出响应。图 6 给出了停车场景的示意图^[32], 其中 P0 至 P5 六个点划分了停车系统可以驾驶和碰撞区域的边界, 而车辆与边界的初始距离 distance to space 以及两个角度变量 gap 和 psi 共同构成了搜索空间的候选解。由于测试的目标是尽可能多地发现该系统中的故障, 因此一个更容易使停车系统产生故障的候选解将被分配更高的适应值。在适应值的计算上, Wegener 等人提出的一种方法是衡量停车过程中汽车轨迹与碰撞边界间的最短距离, 另一种方法则是衡量汽车轨迹与碰撞边界间的面积。这项研究大约模拟了 900 个停车场景, 并从中发现了 25 个会导致碰撞的情况, 其中典型的情形包括车位与汽车距离过远或者汽车初始位置过于接近边界。

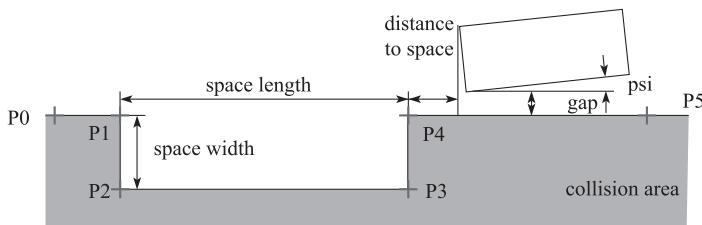


图 6 停车场景的示意图

除了逻辑功能外, 软件系统的正确性还依赖于系统的时间属性, 尤其在实时系统中, 输出产生的过早或过晚都会在某种程度上导致系统的故障。为了确保系统满足指定的时间约束条件, 在执行时间测试中, 测试人员需要生成测试数据来估计系统的最长执行时间 (WCET) 和最短执行时间 (BCET)。然而, 软件系统的复杂性使得这一时间很难估计, 并且执行时间还与当前的硬件环境息息相关。传统的静态分析方法需要对系统的可能路径进行分析, 从而对时间行为进行建模。但这一过程依赖于程序员的辅助信息, 并且容易高估最坏执行时间以及低估最好执行时间。为了对系统的真实执行时间进行更准确的估计, Wegener 等人^[34~36]使用遗传算法来进行测试数据生成。其中, 搜索的适应值

函数是系统的执行时间，而搜索的目标是尽可能地最大化系统的最长执行时间，或最小化系统的最短执行时间。上述研究讨论了一个绘图程序的控制流图，图 7 给出了其最短执行时间和最长执行时间对应的路径。该图中包括循环在内的所有可能的执行路径构成了整个搜索空间，传统的路径覆盖策略或者随机测试都不能很好地找到满足条件的测试数据，而搜索技术在这一问题上则表现得十分有效。此外，Tracey 等人^[37,38]比较了模拟退火和遗传算法在 WCET 测试数据生成上的性能，结论显示遗传算法要比其他算法更加高效。

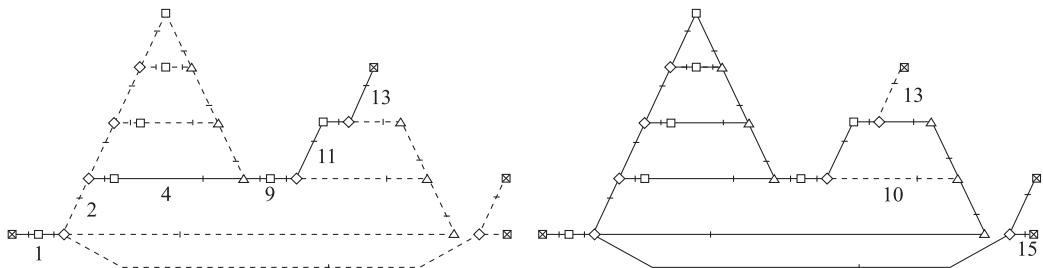


图 7 最短和最长执行时间对应的控制流图

搜索技术也可以为压力测试生成测试数据。压力测试的目标是在持续不断地给待测系统增加压力的情形下，确定系统所能承受的最大压力，相关的测试目标包括最大处理的信息量、最大存储范围以及最大持续时间等。例如，对于实时系统来说，一个触发的事件必须在规定的截止时间内执行完成。因此，在压力测试中，测试人员需要尽可能地构造使系统总是在接近截止时间才能完成任务的测试数据。然而，当系统中存在多种具有不同优先级的周期性或非周期性事件时，测试人员往往难于手工构造合适的测试数据，因此 Briand 等人^[48]尝试使用遗传算法来处理这一问题。假设在测试时间 T 内每个事件 A_i 将被执行 k_i 次，其第 j 次执行的到达时间为 $a_{i,j}$ ，执行的结束时间为 $e_{i,j}$ ，且截止时间为 $d_{i,j}$ ，则测试数据生成的目标是尽可能地减小截止时间和实际结束时间之间的差值 $e_{i,j} - d_{i,j}$ 。这里，遗传算法的每个候选解将包含 $\sum_{i=1}^n [T/\text{min}]$ 个基因，而每个基因都使用二元组 $(A_i, a_{i,j})$ 来表示，其中 n 为事件的数量， min 为最短事件到达间隔。适应值函数将被用于衡量每个候选解的质量，即事件能在截止时间前完成时，候选解将被赋予一个较小的适应值，而一个较大的值则代表一个质量更高的解。Briand 等人^[50]对几个实例进行了研究，发现即使对于理论上可以准确调度的事件，遗传算法也能发现一些会导致系统无法正常执行的情况，因此其建议使用搜索技术来在设计的早期对系统进行验证，并在理论上能正确调度的情况下进一步对关键部分进行测试。类似地，Garousi 等人^[51~53]考虑了分布式实时系统的压力测试，并使用搜索技术来生成达到最大压力的测试数据。此外，Del Grosso 等人^[49]考虑了缓存泄露这一安全相关问题，并使用搜索技术来生成压力测试数据。

在变异测试中，测试人员使用某些变异算子来向原程序中植入一些错误并通过测试来检测这些错误是否能被检出。其中，被植入错误的程序也称作变异体，而当测试用例

集中的某条测试用例检测出原程序和变异体间的不同时，我们就称该测试用例杀死了一个变异体，否则称之为变异体存活。通过分析为原程序设计的测试用例集所能杀死和最终存活的变异体数量，我们就能对原测试用例集的测试充分程度进行评估，如果一个测试用例集能杀死所有植入的变异体，那么这个测试用例集就有能力检测出软件中所有可能潜在的错误^[61]。然而，对每一个变异体都执行一遍测试用例集是一件开销很大的工作，同时还存在一些任何测试用例都无法进行区分的等价变异体，为此人们提出了很多方法来改进变异测试的过程。其中，与搜索技术相关的一项工作是 Adamopoulos 等人^[93]提出的基于遗传算法的变异体和测试用例协同演化策略。在他们的方法中，为了从一系列变异体中选择一个规模较小且较难杀死的变异体子集，首先需要对每个变异体在 0.0 和 1.0 的范围内进行评分，其中一个较高的评分代表该变异体较难杀死。随后，一些随机选择的变异体子集将作为演化的初始候选解，当某个变异体集合 S 中任意一个变异体的评分都不为 1.0 时，则适应值函数为 $f(S) = \sum_{i=1}^n S_i/S$ ，其中 S_i 代表变异体 i 的评分；否则该候选解 S 的适应值为 0。显然，当一个变异体的评分为 1.0 时，该变异体无法被测试用例集中的任一条杀死，因此这样的适应值定义方法能在一定程度上避免 S 中包含等价变异体。类似地，在测试用例选择上，如果一条测试用例能杀死更多的变异体，则该测试用例将被赋予较高的评分，而遗传算法将用于从候选测试用例集中选择一个规模尽可能少且评分尽可能高的测试用例子集。上述两个过程将同时并行地演化，从而不断改进变异体和测试用例集合，最终产生一个较难杀死且不含等价变异体的变异体集合，以及一个能尽可能多地杀死变异体的高质量测试用例集。这里，如果我们分开考虑测试数据和变异体的生成，诸如文献[54~56]等研究尝试使用搜索技术生成能杀死更多变异体的测试数据来提高测试用例集的质量，而 Dominguez-Jimenez 等人^[60]则用搜索技术约简所需变异体的数目来减小变异测试的开销。此外，在变异测试中运用多次变异算子产生的高阶变异体更能模拟真实的软件错误，但传统研究认为构造这样的变异体需要巨大的开销。基于此，Harman 等人^[57~59]提出了基于搜索的技术来为变异测试生成合适的一阶或高阶变异体，Omar 等人^[62]则将类似的方法应用到 Java 程序测试中。

在组合测试中，我们假设系统的故障是由某几个参数间的交互作用所引起的。因此，对于一个包含 n 个参数且每个参数有 l_i 个不同取值的待测系统，测试数据生成的任务即生成一个测试用例集，使得任意 t 个参数间所有可能的组合都在该测试用例集中至少出现一次，而这样的一个测试用例集也被称作 t -way 覆盖表。例如，图 8 给出了一个包含 4 个参数，且每个参数都有 2 个取值 {0, 1} 的 2-way 覆盖表。在覆盖表生成上，当参数和对应取值个数满足某些特定的要求时，一种称为正交表的数学结构可以被快速地构造。在正交表中，任意 t 个参数间的每个可能组合都恰好出现一次，因而该测试用例集具有理论上最小的规模。然而，对于任意的 n 和 l_i ，我们并不知道满足覆盖要求的最小测试用例集大小，因而研究者们提出了很多方法来尝试生成规模尽可能小的覆盖表^[94]。在这一问题上，搜索技术同样

f_1	f_2	f_3	f_4
0	0	0	0
0	1	0	0
1	0	0	0
1	1	1	0
1	1	0	1
0	0	1	1

图 8 一个组合测试
覆盖表

表现出了巨大的潜力。例如，在 Cohen 等人^[63]的研究中，测试人员首先指定一个规模 N 并随机初始化一个 $N \times n$ 的候选数组，这里的适应值定义为该数组中未覆盖的组合个数。随后，搜索技术将不断优化这个候选数组以尽可能地降低其适应值，如果能找到一个适应值为 0 的候选数组，则该数据即为所求覆盖表；否则就递增 N 的值并进行新一轮的搜索。在这一框架下，类似的研究还包括文献[72 ~ 74]等。此外，Bryce 等人^[65,67,68]则使用每次生成一条测试用例的方法构造覆盖表。其中，测试用例是搜索的候选解，而适应值定义为其所能覆盖的新的组合个数，因此搜索技术将用于在每次迭代中生成一条能覆盖最多组合的测试用例，直到覆盖所有组合为止。在这一框架下，类似的研究还包括文献[64, 71, 73, 75]等。从目前的研究来看，搜索技术在覆盖表生成上尽管需要花费比贪心搜索更长的执行时间，但其能极大地减少所需测试用例的数目。此外，由于待测系统中某些参数间的取值组合并不能同时出现，搜索技术同样能在生成时高效地避免生成具有约束条件的测试用例^[69,70]。目前，在组合测试数据的生成领域，搜索技术是主流应用的方法。

除了生成测试数据外，搜索技术同样被成功地应用到其他的软件测试活动中。例如，在集成测试中，测试人员需要依据各构件间的依赖关系来确定构件被集成的顺序。基于层次结构图，传统的集成方法包括自底向上集成（即先集成叶子节点所代表的构件），或自顶向下集成（即从高层逐步向下集成）。对于后一种方法来说，测试人员需要开发桩程序来模拟底层尚未集成模块的对应功能。然而，很多软件构件间会存在依赖关系，从而使得层次结构图中包含一些环，这样传统的集成方法就不能很好地确定合适的集成顺序。因此，集成测试人员需要找到一种合适的构件集成顺序，从而能尽可能地减少所需的桩程序数目，以及尽可能地减少集成所需的步骤或时间，而这两个问题均已知为 NP 完全问题。对于面向对象软件，Hanh 等人^[41]使用遗传算法来解决前一个维度的问题，即最小化所需的桩程序数目。在他们的方法中，某个集成顺序即对应为一个候选解，例如按顺序 {B, D, F, A, H} 集成五个构件，对应的适应值就等于在集成过程中所需的桩程序数。随后，通过求解适应值函数的最小化问题，我们就可以找到优化的集成顺序。Hanh 等人在 6 个真实的实例上比较了包括确定性分析等不同集成方法的性能，验证了搜索技术的有效性。此外，Briand 等人^[42,43]进一步将类间的耦合度纳入集成顺序的考虑因素，并使用遗传算法来寻找面向对象软件的最优集成顺序；da Veiga Cabral 等人^[44]在此基础上使用多目标蚁群算法来对不同的排序标准进行权衡；Colanzi 等人^[45,47]和 DelamareK 等人^[46]则使用搜索算法来解决面向方面的软件的集成顺序问题。

对一个给定的序列按某种准则进行排序是搜索技术应用的典型场景，除了确定集成测试的构件集成顺序外，另一个类似的应用是回归测试中的测试用例集排序^[79,95 ~ 106]。在软件开发过程中，当软件产生更改时，测试人员就需要进行回归测试来确保新的更改并不会对原有软件的功能造成不良的影响。然而，随着软件功能的不断开发，回归测试用例集的规模也将逐步增长，而可供测试的资源往往又十分有限，因此如何从中选择合适测试用例进行执行，以及如何安排测试用例的执行顺序来更快地检测出软件中的故障，是回归测试中需要考虑的关键问题。对于测试用例排序来说，给定一个测试用例集 T 以

及评价函数 f , 我们期望找到 T 的某个执行顺序 P , 使得对任意的 $P' \neq P$ 都有 $f(P) \geq f(P')$ 。例如, 表 3 给出了四条测试用例的语句覆盖情况^[79], 当按 $\langle A, B, C, D \rangle$ 的顺序来执行时, 测试完前两条测试用例后仅覆盖了其中 6 条语句。而如果我们按 $\langle C, D, A, B \rangle$ 的顺序来执行, 测试完前两条测试后所有 8 条语句都得到了覆盖。在排序问题上, 由于某个执行顺序 P 可以很容易地编码为适合搜索技术应用的形式 (例如字符串), 排序目标也可以很容易地转换为适应值函数, 因此搜索技术非常适合于解决排序问题。目前, 基于覆盖率、测试执行时间或者测试需求等排序场景中都有搜索技术的应用。此外, 在回归测试中, 如何从测试用例集中选择一个子集, 从而对软件更改的部分进行测试也是搜索技术应用的一个场景^[76~78], 另一个回归测试的应用场景则是对已有测试用例中冗余的部分进行约简, 从而减少所需测试用例的规模^[82]。在排序算法的研究中, Li 等人^[79]在 2007 年首次比较了贪心算法、额外贪心算法、遗传算法等五种不同的搜索算法在测试用例优先排序方面的效率差异, 随后又继续开展了基于多目标的启发式搜索算法在测试用例优先排序上的研究^[104], 使用了基于非支配排序的遗传算法改进算法 (Non-dominated Sorting Genetic Algorithm II, NSGA-II), 并结合异构并行技术 GPGPU (General-Purpose computation on GPU) 技术, 提出了粗粒度和细粒度两种并行计算策略, 极大地提高了搜索算法的效率, 为测试用例优先排序技术在产业界的实际应用提供了支持。

表 3 测试用例的语句覆盖情况

测试用例	语句覆盖							
	1	2	3	4	5	6	7	8
A	×	×	×			×	×	×
B	×	×	×				×	×
C	×	×	×	×				
D					×	×	×	×

此外, 测试用例的预期输出也是软件测试中的一个关键问题。McMinn 曾提出一种基于搜索的测试预期生成方法来对浮点数进行测试^[81]。在这一方法中, 为了检测原始程序输出的正确性, 可以应用一种类似 N -版本编程的方法来生成待测程序的另一个版本, 进而通过比较两者间输出的差异来判断测试结果是否满足预期的测试输出。然而, 大多数情况下测试用例的预期输出还是需要人工来进行确定, 但搜索技术仍可以在这一问题上减少测试人员的人工成本开销。从数量上看, 合适的适应值函数可以使搜索在最大化测试目标的前提下最小化所需执行的测试用例数目^[108,109], 从而减少需要人工评估的测试结果。而从质量上看, 搜索技术也可以通过增强测试场景的可辨别性来帮助测试人员更加容易地判断所执行的测试数据是否通过测试^[110]。

总之, 面对规模日益庞大和复杂的软件, 基于搜索的软件测试避免了传统测试和分析方法的不足, 能在尽可能降低测试成本的前提下尽可能高效地完成不同的测试任务, 是软件测试领域中一种高效实用的新方法。

2.2.2 程序错误自动修复与错误定位

程序错误自动修复技术综合利用错误定位、修复规则以及补丁验证技术对可复现的

错误进行修复。20世纪80年代软件生产自动化被广泛关注，但软件维护阶段的自动化技术受阻于程序逻辑的复杂性一直没能达到工业级应用要求。

随着21世纪软件资源挖掘和搜索技术的流行，自2009年以来以测试为基础的逻辑无关或弱逻辑的软件错误自动修复技术开始出现工业级应用的前景。（图9给出了软件错误自动修复技术框架）。随后，研究者发现了程序错误自动修复中量变和质变现象^[112]，利用面向自动修复的错误定位^[113]，提高了定位的准确度，推动了软件维护自动化进程。2012年，针对百万行软件错误修复代价高昂的问题，研究者从软件自动修复三阶段（即错误定位、补丁生成和补丁验证阶段）入手，建立了大型开源软件历史错误分析数据集和错误定位方法有效性分析比较平台，进行错误定位有效性分析，形成错误定位方法的改进完善实验床^[114,115]；分析程序错误修复过程中可能的复用信息和过程，在补丁验证中引入基于编译现场缓存的弱编译技术，提升了验证过程中占据主要耗时的编译的效率^[114,115]；在补丁验证的回归测试用例集合中，引入测试用例动态序概念，加快了无效补丁的发现效率，降低了验证时间^[116]。

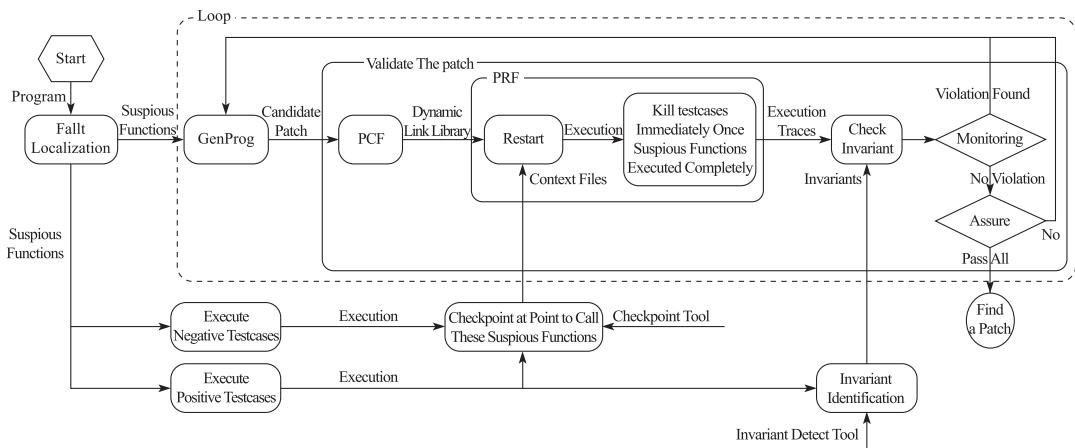


图9 软件错误自动修复技术框架

错误定位通过对程序失效或者异常行为进行分析，寻找导致失效或者异常的根源所在，即定位到错误在程序中的位置。软件错误定位技术的先进性主要体现在错误定位的准确性和实用性。现有的错误定位技术主要关注于两个方面：一是如何找出尽量小且包含缺陷语句的可疑语句集合；二是如何赋予缺陷语句更高的可疑值，提升缺陷语句在可疑语句排序表的位置。主要定位技术涉及基于切片、基于统计、基于程序状态、基于机器学习等四个方面。

研究者通过构造面向错误定位的轻量级近似动态后向切片算法，提出了程序逻辑与统计相结合的软件错误定位方法SSFL，构造出基于程序逻辑和差异信息的度量公式，通过控制流依赖和数据流依赖分析，融合差异性比对获得的统计信息，提高了定位准确度^[117,118]。图9即为软件错误自动修复技术框架示意图。

2.2.3 需求分析

需求分析作为软件生命周期的第一个阶段，并贯穿于整个软件生命周期。随着工程规模的扩大，需求工程的重要性越来越突出，因为在初期阶段的频繁和细微的变化都会对后期带来大量的修改工作。如何在工程初期做好需求分析将直接决定整个工程的成败。

事实上，需求工程常被划分为需求开发和需求管理两部分贯穿于软件工程的整个生命周期之中。随着软件规模的日益增长和复杂性增加，每个客户都有自己的优先需求目标（见图 10），不同的需求之间又通常会存在利益冲突。此时，问题就是一个典型的 NP 问题。为了在现有资源中权衡满足每个用户的需求做出最优的选择，需求工程师们选择启发式优化算法解决复杂的、多目标的、条件约束的需求分析问题。

在需求工程中，以下一版本问题为例（Next Release Problem, NRP）^[120]，通常采用启发式算法解决探索成本消耗和利润之间的平衡关系。在 NRP 中，唯一的目标是挑选出最优的需求解，在满足约束条件下使客户满意度达到最大，最大化客户利润的同时最小化成本花销。优化结果能给决策者在需求选择的问题上提供充足的理论支撑和证据。Bagnall 等人提出 NRP 模型解决单目标优化问题。假定有一个软件项目，存在客户群 $C = \{c_1, \dots, c_m\}$ ；提出了一系列软件需求表示为 $R = \{r_1, \dots, r_n\}$ ；相对应的需求项目消耗 $\text{Cost} = \{\text{cost}_1, \dots, \text{cost}_n\}$ ；每个客户对公司的重要性不同，所占权值表示为 $\text{Weight}_i = \{w_1, \dots, w_m\}$ ($w_j \in [0, 1]$ 且 $\sum_{j=0}^m w_j = 1$)；将评价需求重要性的分数值表示为 $\text{Score}_i = \sum_{j=0}^m w_i * \text{value}(r_i, c_i)$ ；决策向量 $x = \{x_1, \dots, x_n\} \in \{0, 1\}$ ，当 $x_i = 1$ ，表示需求 i 将选入下一优化版本中，否则 $x_i = 0$ 。

随着解决需求的复杂化，传统的单目标优化问题逐步发展成为多目标优化问题（Multi-Objective Next Release Problem, MONRP）^[121, 122]。基于多目标的帕累托最优化策略（Multi-objective Pareto Optimal），通过启发式算法挑选出不受其他变量支配的非支配子集（non-domination population），每一个非支配集合表示了一种可能的资源分配，并且在已有资源限制条件下使客户满意度达到最优。需求目标的适应度计算公式如下^[123]：

$$\text{Maximize} \sum_{i=0}^n \text{score}_i * x_i \quad \text{Minimize} \sum_{i=0}^n \text{cost}_i * x_i$$

Zhang 等人^[121]使用 Motorola 需求数据集合，采用了 NSGA-II 算法进行了多目标需求优化的实验。如图 11 中所示的帕累托前沿（Pareto Front），每个圆圈表示当需求消耗不变时所能构成的最优解决方案。实验证明随着客户满意度的增加，执行客户需求的消耗值也随之增加。

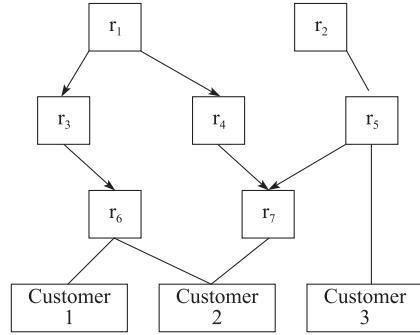


图 10 客户需求结构

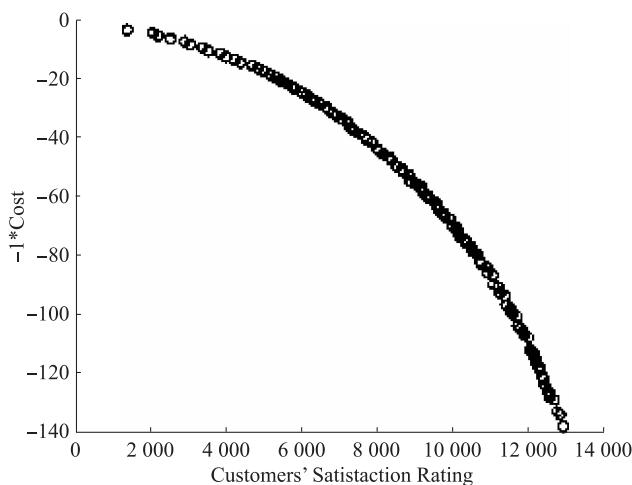


图 11 基于 Motorola 数据集的实验结果^[121]

随之，越来越多的基于多目标的优化算法应用到需求分析中^[124~132]。Tonella 等人提出了一种交互式的遗传算法优化需求顺序^[125,126]，Kumari 等人采用了精英量子演化算法 (Quantum-inspired Elitist Multi-objective Evolutionary Algorithm, QEMEA)^[127]，Jifeng 等人使用了骨架分析与启发式算法^[128]等。此外，还包括多目标的混合式量子差异进化算法 (Multi-objective Quantum-inspired Hybrid Differential Evolution, MQHDE)^[129]、双归档算法 (Two-Archive algorithm)^[131]、聚类的方法^[132]等应用到多目标到需求选择优化问题。Antônio Mauricio Pitangueira 等人^[133]分析了这一领域的 30 多篇文献，系统地评价了不同的优化算法，指出由于不同的方法具有不同的优缺点，研究者们可以根据问题的特征选择较优的算法解决需求优化问题。

在众多的算法研究中，基于搜索的需求工程同时还研究需求分析中的公平性分析^[134]和敏感性分析^[136]等。

(1) 公平性分析 (Fairness Analysis)

公平性分析的动机在于平衡客户的需求执行度，尽可能满足每个人的首要需求。然而，不同的人对待公平的定义有不同的标准，Finkelstein 等人^[135]首次在不同的公平定义中寻求公平在不同人标准的平衡，采用了基于搜索的技术揭露了客户需求间的依赖关系。非支配排序的遗传算法 (Non-dominated Sorting Genetic Algorithm-II, NSGA-II) 是目前最流行的多目标进化算法之一，它降低了非劣排序遗传算法的复杂性，具有运行速度快，解集的收敛性好的优点，成为其他多目标优化算法性能的基准。在需求工程中的公平性分析问题中可以看成是多目标优化问题。

Finkelstein 等人分别使用不同的数据集合 (Motorola data set, Greer data set, Random data set) 进行测试^[136]，实验发现执行的需求的数量与公平性成反比。当完成的需求越多，客户之间的公平性就逐渐降低。总的来说，基于搜索技术将逐步应用在现实生活中，揭示出暗藏在数据集中的潜在关系。但是，需求工程作为新兴学科方向，还有较大的探索空间等着我们去发掘。

(2) 敏感性分析 (Sensitivity Analysis)

敏感性分析是研究与分析一个系统（或模型）的状态或输出变化对系统参数或周围条件变化的敏感程度的方法。在最优化方法中经常利用灵敏度分析来研究原始数据不准确或发生变化时最优解的稳定性。通常，当一个系统变得越来越复杂，就越来越难获取输入输出之间的关系。灵敏度分析可以基于参数输入和观察到的输出结果分析出它们之间的关系，通过灵敏度分析还可以决定哪些参数对系统或模型有较大的影响^[136,137]。

2.2.4 软件设计

在软件开发流程中，软件设计是从需求分析向软件具体设计进行转化的一个步骤。通常来说，设计人员首先需要依据规格说明来确定软件的一些基本结构，随后通过进一步分析并进行相应地修改，设计人员就可以确定软件的最终设计。除此之外，对软件质量进行增强和预测的相关研究，也通常被视为软件设计研究的组成部分。近年来，基于搜索的软件设计也发展得较为迅速，从软件的高层体系结构设计到软件聚类和软件重构，搜索技术在很多方面都取得了成功的应用。Raiha^[138]曾对搜索技术在软件设计领域的应用进行了总结，尤其强调了在不同问题下的编码和适应值函数的选择问题，并对已有方法的性能进行了评价和讨论。

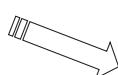
软件的体系结构设计是软件设计的核心，这一工作需要设计人员有丰富的经验来从高层的需求中确定具体的细节设计。例如，在面向对象的软件设计中，设计人员首先需要从使用案例中抽取信息来确定每个类的方法和属性，并在此基础上确定接口和继承关系，从而得到类似 UML 类图的表现形式。其中，使用案例是用户为了达到某种目的而执行的一系列时序步骤，描述了用户与系统间的交互场景。使用案例通常以文本形式来记录相关操作和对应数据，例如“系统扣除某账户的若干余额”就描述了系统中的某一具体行为。从使用案例中可以确定系统所包含的行为和数据，然后通过为每个类分配合适的方法和属性，就可以得到具体的一种软件设计，而所有可能的方法和属性的集合就构成了面向对象设计问题的求解空间。图 12 给出了面向对象软件底层设计的流程^[138]。

Use Case:

```

Actor requests Action1
System performs Action1 with Datum1
Actor requests Action2
System performs Action2 with Datum2

```

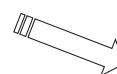


Method / Attribute derivation:

```

Action1 ▶ Method1
Datum1 ▶ Attribute1
Action2 ▶ Method2
Datum2 ▶ Attribute2

```



Class:

Attribute1	Attribute2
Method1	Method2

图 12 面向对象软件底层设计的流程

由设计人员人工来确定最优的面向对象软件设计往往是很困难的，而搜索技术不仅能产生与人工设计质量相当的结果，并且能在一定程度上生成许多人工设计未曾考虑过的更优结果，因而是这一领域的一种较有潜力的方法。当应用搜索技术时，一个候选类图解可以被编码为一个整形数组，其中某一位代表该系统中的一种方法或属性，而对应的整数值代表该方法或属性被分配到的类。这里的适应值函数通常是基于方法或属性的内聚度（cohesion）或耦合度（coupling）来进行设计。例如，对于类 C 来说，其方法的内聚度（COM）定义为 $f(c) = 1/(|Ac| |Mc|) * \sum \Delta_{ij}$ ，其中 Ac 和 Mc 分别代表了类中属性和方法的个数，且当方法 i 使用了属性 j 时有 $\Delta_{ij} = 1$ ，否则 $\Delta_{ij} = 0$ 。此外，由于对软件设计的某种性质进行优化往往会削弱其他性质，因此这一领域所使用的方法大多是基于多目标优化的搜索策略。

例如，Simons 和 Parmee^[139~141]使用多目标优化的遗传算法来从使用案例设计面向对象软件，他们所使用的适应值函数包括三个部分：类的内聚度、类间的耦合度以及类的数量。类似地，O’Keeffe 和 O’Cinneide^[142,143]使用模拟退火方法来求解这一问题，并对一系列的适应值函数进行了比较。Bowman 等人^[144]则针对已有软件设计进行优化的问题，提出了基于多目标遗传算法的决策支持系统来优化类中方法和属性的分配。此外，上述设计方法主要考虑的是面向对象的底层设计，搜索技术同样可以进一步用于软件体系结构或设计模式的实现。为了提高软件的可重用性，Amoui 等人^[145]使用遗传算法来寻找最佳的模式实现顺序，从而构建设计模式的最优转化步骤。而通过将上述研究工作与设计模式的相关研究进行结合，Raiha 等人^[146,147]提出了基于遗传算法的搜索技术来自动化地对包含多种设计模式的软件体系结构进行合成。

在面向对象软件设计中，搜索技术的另一个热门应用是软件行为模型的设计。高可信软件通常需要依赖自动系统对外界计算资源、组件和物理环境的变化做出合适的响应，然而自动系统的行为在部署前往往很难预测，因此自动化地生成软件行为模型能极大地帮助软件工程师来理解系统的行为。在这一问题上，Goldsby 等人^[148~150]开发了基于搜索的软件行为模型生成工具 Avida-MDE。这一演化工具使用描述系统行为的 UML 状态图作为候选解，并通过选择和变异操作来不断演化。其中，每个个体模型的适应值将依据开发人员所定义的一个任务集合来评估，能成功执行的每个任务都将增加候选模型的适应度。类似地，Lucas 和 Reynolds^[151]同样使用搜索技术来对确定性状态机进行学习，从而更好地对软件系统的状态标签进行分配。

搜索技术同样可以用于面向服务的软件设计。通过将应用程序的功能作为服务发送给最终用户或者其他服务，面向服务的体系结构可以非常灵活地结合不同的服务以提供一些复杂的功能。其中，服务的后绑定机制是面向服务软件的核心，这就使得面向服务系统可在运行时选择所需的具体服务。例如，在图 13 所示的例子中^[152]，圆形所代表的是设计中需要使用的服务（抽象服务），而正方形所代表的是具体实现该功能的可选服务（具体服务）。与同一抽象服务相关的具体服务在功能上是等价的，因此人们通常按照服务质量（Quality of Service, QoS）这一非功能性标准来在其中进行选择，例如考虑服务的价格、响应时间、可用性以及名声和用户满意度等指标。

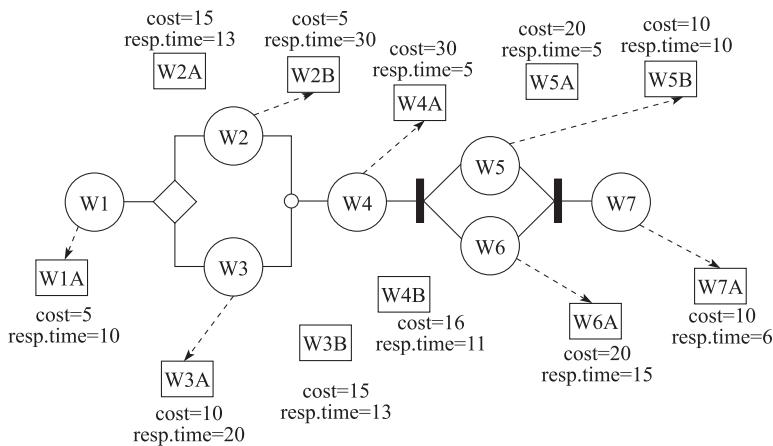


图 13 抽象服务与具体服务间的绑定示例

给定需要的抽象服务和可选的具体服务，如何进行对应的绑定以使得系统在满足一定约束条件的前提下尽可能地优化某些特定目标，是面向服务软件中的关键问题。这一问题也称作 QoS 的服务组合问题，而其最优解的寻找已被证明是 NP-hard 的。线性整数规划可以用于寻找该问题的解，但其求解时间并不能满足当前系统的快速构建需求。基于此，Canfora 等人^[152]提出了一种优化服务选择的遗传算法。在他们的方法中，一个组合服务的候选解 S 用包含 n 个服务的整型数组 $[s_1, s_2, \dots, s_n]$ 来表示，其中 s_i 代表与第 i 个抽象服务进行绑定的具体服务。由于需要考虑的 QoS 指标包括时间、开销、可用性和可靠性，因此适应值函数定义为：

$$f = (w_1 \text{Cost} + w_2 \text{Time})/w_3 \text{Availability} + w_4 \text{Reliability} + w_5 D$$

其中， w 代表了不同指标的权重， D 则代表了候选解偏离满足约束条件的距离，从而对无效的候选解进行惩罚。虽然上述适应值的计算看起来较为简单，但其中 Cost 和 Time 尤其是 Availability 和 Reliability 等指标的衡量往往需要复杂的计算，并且需要针对具体系统进行具体设计。显然，一个较小的适应值代表该候选组合的质量越高。除了考虑对新的服务进行组合优化外，Canfora 等人^[153]随后继续将遗传算法用于服务运行时的重绑定问题，从而使得原先确定的服务在服务质量与预测值产生偏差时进行服务的重新组合。

此外，Jaeger 和 Muhl^[154]考虑了在不同的 QoS 特征下的服务组合问题，并使用遗传算法来进行优化。Khoshgoftaar 等人^[155, 156]则使用多目标优化的搜索技术来平衡不同的 QoS 服务。在算法设计上，通过对遗传算法的编码、初始化方法和演化机制进行改进，Zhang 等人^[157~160]提高了遗传算法在优化服务选择上的性能，从而更高效地解决了约束条件下的 QoS 服务选择问题。而与传统的多目标优化不同，Cao 等人^[161]指出很多商业流程中开销是需要首要考虑的因素，并使用单目标优化的遗传算法来优化服务选择流程。

软件设计中的构件选择问题同样可以用搜索技术来优化。由于软件开发大多是迭代进行的，基于构件的方法可以不断地在当前状态下为软件加入新的构件以使得软件不断演化。然而，为软件的下一个版本选择构件并不是一项简单的工作，项目管理人员通常需要对开发开销、用户期望、开发时间、预期回报等指标进行综合考虑来做出最合理的

选择。同时，现实中这些指标间往往存在一定的权衡，这就使得最优选择的确定变得十分困难。此外，对于选择好的一些构件，如何安排它们的实现顺序以最大化项目收益也需要谨慎考虑。类似于服务选择问题，假设当前共有 n 个构件 x_1, \dots, x_n 可供选择，可用的开发资源为 K ，其中开发每个构件所需的成本为 c_i ，且对应的预期的收益为 w_i ，则构件选择问题就等同于背包问题，即选择某 m 个构件以使得在满足约束条件的前提下最大化收益。由于这一问题是已知为 NP-hard 的，为了寻找尽可能好的解，Baker 等人^[95]、Vijayalakshmi 等人^[162]以及 Yang 等人^[163]分别使用模拟退火和遗传算法来对其进行优化。传统的运筹学方法在这一问题上同样被广泛使用，例如 Desnos 等人^[164]曾尝试将回溯法和分支限界法进行结合。此外，Cortellessa 等人^[165]提出了一个旨在最小化系统构件成本的构件选择框架，Kuperberg 等人^[166]提出了一个基于遗传规划的重工程参数行为模型来进行黑盒构件性能的预测。

在软件设计领域，Feldt^[167~169]还尝试对容错进行考虑，其使用遗传规划方法来对软件进行演化以得到软件的多个版本，从而帮助开发人员在 N -版本容错中降低开发不同软件版本的开销，同时提高整个软件系统的鲁棒性。在其他的软件设计问题上，Barlas 和 Rl-Fakih^[170]使用遗传算法来优化多客户端和多服务器系统中的应用交付问题；Chardigny 等人^[171]使用搜索技术来解决面向对象体系结构中基于组件的抽取问题；Sharma 和 Jalote^[172]使用搜索技术来优化所部署软件构件的性能。

2.2.5 软件重构与维护

软件维护费用占软件开发总费用的 40% 以上。程序开发者通过软件重构、程序分析等手段提高软件的灵活性、可重用性等方面进而降低软件维护开销。基于搜索的软件工程利用搜索算法寻找有价值的软件重构方式或程序片组合模式，进而提高软件维护过程的效率，最终达到自动化或半自动化软件维护的目的。利用 SBSE 进行软件维护通常包括三个步骤：1) 建模软件维护问题；2) 设定合适的目标函数；3) 选择合适的搜索算法。基于搜索的软件工程在软件重构与维护问题上的主要研究内容包括软件重构与程序分析两部分。

软件重构是 SBSE 在软件维护中的主要研究方向。早期的基于搜索的软件重构技术集中在利用搜索技术提高程序执行效率、减少程序规模，该过程主要是通过启发式算法搜索并优化程序中的循环语句、冗余语句等，寻找更高效的代码表现形式^[173~175]。

随着面向对象语言的成熟，研究者尝试结合面向对象语言的特性利用基于搜索的软件重构方式进行自动化或半自动化的软件重构工作^[176~178]。在自动化软件重构过程中，研究者首先对软件重构问题进行建模，如从代码级别、方法级别甚至是模块级别^[179]进行问题建模，寻找符合搜索算法的域编码方式，并在此阶段确定不同域的重构规则，如降低域的继承层次、增加子类等规则。除针对软件源代码进行建模外，也有研究者对软件中说明性语言（declarative languages）进行建模，研究软件说明性语言的自动化重构^[180,181]。其次 SBSE 需要设定合适的目标函数（fitness function）。在目前的研究中，软件重构目标函数以 QMOOD 度量为主，QMOOD 度量从软件的灵活性、可重用性、可理解

性等方面评价软件重构结果的优劣。也有研究者从其他角度，如重构软件的可测试性等方面，进行软件重构结果的度量^[182,183]。由于软件重构结果可以从多个角度进行度量，因此研究者尝试联合多个目标函数进行多目标的软件重构优化^[184]。最后 SBSE 需要选择合适的搜索算法，大量的搜索算法被尝试用于进行软件重构，并且研究者通过经验学习的方式在多个数据集上结合多种软件重构规则来对比不同搜索算法在软件重构的可用性^[185,186]，所对比的搜索算法如遗传算法、模拟退火算法、爬山算法等。

由于自动化软件重构可能会产生大量无意义的重构模式，研究者又尝试引入人机交互进行半自动化软件重构，通过对重构中间结果的人工干预，使得最终的重构软件符合预期。常见的方式如交互式遗传算法（interactive genetic algorithm）、可视化的 Pareto 最优曲线的引入^[178,184]等。

SBSE 在软件维护中的另一个研究方向是程序分析。程序分析这里指依赖分析（dependence analysis）与概念分配（concept assignment）。自动化程序分析能够帮助软件维护人员快速理解整个待维护的软件系统。在依赖分析方面，研究者利用遗传算法、贪心算法等搜索能够覆盖所有程序功能点的程序片组合，帮助维护人员进行程序依赖结构分析^[187]。在概念分配方面，研究者利用搜索算法寻找可能的程序片组合方式，挖掘便于理解的高层概念^[188]。

2.2.6 软件项目开发管理

软件项目开发管理是整个软件开发过程的基础，它通过分析软件项目规模、开发人员专业度等因素来辅助项目经理对软件项目的人员、开发任务进行合理分配，达到减少软件开发时间、减少人员时间碎片的目的。SBSE 把软件项目开发管理建模为对软件项目中多种资源的组合优化的问题，利用搜索算法进行高效的资源分配和项目评估。软件项目开发管理涵盖软件项目开发中的时间管理、花销管理、质量管理、人力资源管理、风险管理等，它可以分为软件项目资源优化和软件评估两方面。

在软件项目资源优化方面，研究者利用分散搜索（scatter search）、遗传算法并结合软件项目中已知的可更新资源、人员等级等进行软件项目中的人员分配优化和工作模块分配优化，并研究如何根据人员的数量提出不同的分配意见^[189,190]。另外，人员交流情况、人员专业度等特征也可作为影响软件项目优化的因素^[191,192]。除进行单目标的资源优化外，软件项目中往往需要同时对诸如项目完成时间、项目花销等多目标进行优化。研究者提出利用多目标搜索的方式完成该任务，如利用多目标遗传算法、NSGA-II, SPEA2 等演化算法^[193,194]。也有研究者从迭代的角度对多个优化目标进行迭代交叉优化，或以基于事件的方式在某个资源发生变化时对其相对应的目标进行优化^[189,195]。考虑到多目标优化结果的多样性，研究者利用可视化的方式让决策者从多个优化方案中进行选择，如 Pareto 最优或对每一个目标利用仿真模拟器生成其相对应的结果供决策者选择^[191,196]。

软件评估是软件项目开发管理中的重要活动，利用软件评估结果可以进行任务分配、人员分配等任务，典型的软件评估工作如软件规模评估、软件开发开销评估等。研究者

对已有的机器学习软件评估方法与基于搜索的方法进行对比，证明后者能够完成软件评估任务，并可能获得更好的准确性，但是基于搜索的方法在程序设置和运行开销上优势并不明显^[197,8]。因此研究者提出利用混合策略把已有的软件评估方法和启发式算法相结合，进行联合的软件评估，提高评估的准确性，例如把基于类比（Analogy-Based Estimation, ABE）的评估方式与粒子群算法相结合^[198]。大量的搜索算法（如遗传算法、爬山算法、模拟退火算法等）被用于进行软件评估工作，同时有研究者对不同算法、不同项目规模、不同建模方式进行对比，评估各种因素在软件评估方面的优劣^[199,200]。除了常见的软件规模评估、软件开销评估外，也有研究者利用 SBSE 进行软件复杂度、耦合度等影响其质量的因素的评估^[201]。

在软件项目开发管理中，与上述利用软件项目资源优化结果和软件评估结果来间接辅助项目经理决策不同，SBSE 也可以直接基于已有的项目决策数据，搜索并生成优秀的项目决策组合方案。研究者可以根据项目经理对已有项目决策的标注结果，结合软件项目模拟器（software project simulator）的模拟数据，利用搜索算法搜索出大量“GOOD”决策，或直接把搜索算法应用到范例推理系统中（Case-Based Reasoning system, CBR）^[202,203]。

2.3 基于搜索的软件工程在工业界的应用

戴姆勒－克莱斯勒公司（DaimlerChrysler AG）较早就对基于搜索的软件工程的工业运用进行了尝试。为了满足单元测试的需要，J. Wegener 等人设计并搭建了完整的演化测试环境^[204]，其核心思想是利用遗传算法完成测试数据的自动生成。文中将不同的测试准则分为四类，并提出了针对性的适应度函数。A. Windisch 等人^[205]于 2007 年提出运用完全学习粒子群算法（Comprehensive Learning Particle Swarm Optimization, CL-PSO）来代替遗传算法。实验比较两种算法在 13 个工业程序上的测试数据生成效率，发现了 CL-PSO 在各个被测程序上的测试用例生成效率上不劣于 GA，并且在更复杂的被测程序上要明显优于 GA 算法。Harman 等人分析了输入域约减对基于搜索的测试数据生成的影响^[206]，通过实验证明了随机测试、爬山算法以及演化测试对搜索空间的敏感程度，得出了随机测试不受输入域约减的影响以及演化测试较爬山算法受影响程度更大的结论。在以上研究与工业实践的基础上，O. Buhler 和 J. Wegener 通过模拟具体场景的方法对演化测试的应用范围做了进一步的延伸^[207]，设计了针对两个特定的系统（自动停车系统和辅助刹车系统）的功能性演化测试方法。

微软公司的 PEX 是工业应用广泛的软件测试工具。PEX 与其他许多工业级的软件测试工具相同，也使用了符号执行（Symbolic Execution）技术^[208]。一方面 T. Xie 等人^[209]将基于搜索的软件工程技术与 PEX 中的符号执行技术相结合，通过利用适应度函数引导的搜索策略改进传统的符号执行的路径搜索策略，使得新的测试方法能够更早地达到较高的覆盖率。另一方面，K. Lakhotla^[210]将基于搜索的软件测试引入动态符号执行（Dynamic Symbolic Execution），分别利用 AVM（Alternating Variable Method）以及 ES（Evolution Strategies）两

种搜索算法来改进 PEX 对浮点数约束处理能力的不足。实验比较了 PEX 以及两个应用搜索算法的 PEX，结果显示后者确实能提高 PEX 对包含浮点数计算的程序的覆盖率，但是大量的适应度函数计算也带来了更多的执行时间开销。

谷歌公司使用大量的并行设备来执行大规模的回归测试。但测试用例集过大以及高频率的代码变化都限制了回归测试的效率。因此，谷歌希望在代码被提交到并行平台上进行大规模的回归测试前，开发者就能够在本地对部分模块先进行小规模的测试。针对这一需求，S. Yoo 等人于 2011 年利用回归测试优化技术设计了新的测试框架^[211]。这一框架包含本地测试以及全局测试两部分。通过使用 Two-Archive 多目标演化算法（Multi-Objective Evolutionary Algorithm, MOEA）对四个优化目标：依赖覆盖、故障历史、执行时间以及误报测试用例过滤进行优化，选择出测试用例子集用于提交前的本地测试。提交前的测试优化有以下三个优点：1) 避免了开发者人工确定测试用例与被修改的模块之间是否关联的开销；2) 能够给予开发者更早的测试反馈信息；更早地对错误进行修复提高了整个测试过程效率。基于保持已有测试框架完整性的考虑，在并行设备上进行的全局测试需执行所有与改动关联的测试用例。这样保证了测试用例集能够发挥全部错误检测能力。实验结果显示，优化的测试用例子集相比较测试用例全集可以减少 33% ~ 82% 的测试时间。

IBM 公司将多目标优化算法应用于大规模测试用例最小化中，并使用 GPGPU 提升了算法效率。S. Yoo 等^[211]指出频繁的适应度计算耗费了大量的时间，这使得许多基于搜索软件工程的研究不能直接应用到工业上。然而，GPGPU 并行构架的出现为这一问题提供了可行的途径。文章提出了一种基于搜索的优化算法来解决测试用例最小化问题，并且用 GPU 加速了该算法。除了实验室级的标准测试程序，工业界的程序同样被用来评估文章提出的方法。实验表明，通过使用 GPGPU 技术能够获得超过 25 倍的加速比。

IBM 现有的测试用例选择方案能够从测试用例集中选择出多个测试用例子集。但是实验发现，有一些测试用例频繁出现在各个测试用例子集中。为了避免每次回归测试频繁地执行同样的测试用例，每一次回归测试的时候不选择那些之前已经执行过的测试用例。这种策略会使得测试用例池的规模慢慢减小，同样降低了测试用例子集的代码覆盖率。S. YOO 等人^[212]提出了一种测试用例约简和生成相结合的方法，在从测试用例集中选出测试用例子集后，重新生成一些测试用例来保证测试用例集的完整性。针对测试用例的再生成，文献[212]提出了几种策略：爬山算法和分布估计算法（Estimation of Distribution Algorithm, EDA）。实验将结合测试用例生成的方法与原有方法对比，通过记录回归测试的每次迭代选择出的测试用例子集能够达到的最大代码覆盖率，验证了文章提出的方法能够在减少测试用例重用的基础上保持测试用例子集的代码覆盖率。

摩托罗拉公司运用基于搜索的软件工程技术来解决软件最优发布问题。P. Baker 等人^[213]在文中指出，随着基于组件的软件开发的广泛应用，管理者更多地面临困难的决策：在综合考虑众多影响因素时，如何从大量的候选组件中选择出合适组件以及确定组件的优先级。这些因素可能包括：获取开销、用户意愿、开发时间、预期回报、组件间的依赖关系和不同用户的优先级。通过将这一决策问题转化为最优化 0-1 背包问题

(optimization 0-1 knapsack problem)，文章提出使用模拟退火算法以及贪心算法的解决方法。实验对比以上两种算法以及专家人工评估的方法在一个包含 40 个组件的软件库上的表现，结论显示两种自动的方法都优于人工的方法，并且模拟退火算法的效率以及稳定性都优于贪心算法。

爱立信公司运用基于搜索的软件工程技术解决需求分析和优化的问题。Y. Zhang 等人指出虽然需求提取以及分析是软件开发过程中的第一个行为，但是在实际开发过程中需求却经常发生变化^[214]。因此，文章提出了现在/未来重要性分析（Today/Future Importance Analysis, T/FIA）。需求的变化可以分为可预知的以及不可预知的两类。文章 [214] 利用可预知的需求变化以及不可预知中已知很可能发生的需求变化来构造未来的需求。针对当前需求、未来需求以及开销三个目标，使用 NSGA-II 算法对原始需求集进行优化。实验数据使用了来自爱立信公司 14 个不同软件测试子部门对同一测试管理软件的需求。结果显示 NSGA-II 算法产生了具有良好分布的 Pareto Front 最优解集。

尽管基于搜索的软件测试已经有了大量的实际应用，但是依然缺乏相应的开放工具。K. Lakhotia 等人^[215]设计并实现了针对 C 程序的基于搜索的开源测试数据生成工具 AUSTIN (AUgmented Search-based TestINg)。AUSTIN 可以生成大多数类型的测试数据，除了字符串、空指针、函数指针和结构体。AUSTIN 支持三种搜索算法：随机搜索、爬山法和结合符号执行的爬山法。实验选择 8 个 C 函数作为评估的基准程序。这 8 个程序选自汽车的自动控制系统，包括自适应车前灯控制、门锁控制以及电动窗控制系统。通过与先进的演化测试工具 ETF 进行比较，AUSTIN 在生成满足分支覆盖准则的测试数据上具有更高的效率。

3 国内研究进展

在国内，SBSE 的研究发展也很有起色。目前，已经有很多高等院校、研究所的教师和学生从事该方向的研究工作，并组织了中国基于搜索的软件工程研讨会（Chinese Search Based Software Engineering Workshop）[⊖]。第一届于 2012 年 7 月在北京化工大学召开，第二届于 2013 年 6 月在大连理工大学召开，第三届于 2014 年 7 月 4 日在中国矿业大学召开。研讨会把国内从事 SBSE 相关的学者组织在一起，逐步发展并形成了一定规模的研究人员社区。

国内研究人员紧跟 SBSE 研究热点，很多研究成果都发表在重要国际会议和国际期刊上，在上一章节的综述中已经包括。本章我们仅收集了《软件学报》、《计算机学报》、《计算机研究与发展》和《电子学报》近五年刊发的与 SBSE 相关的文献，历年相关的文献发表情况如图 14 所示。可以看出，从 2009 年至今，每年发表的有关 SBSE 的文献数量在稳步上升，而软件测试在其中占据了很大的比重。

⊖ <http://www.csbse.org>

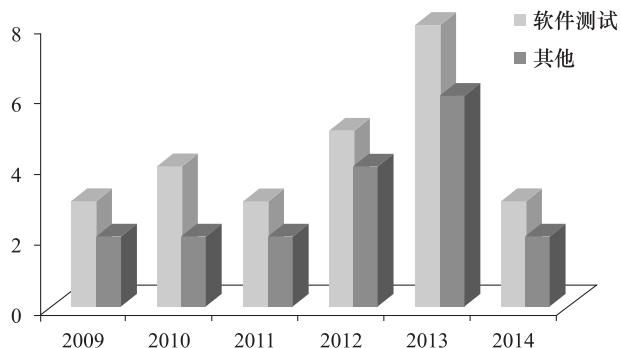


图 14 SBSE 方向论文发表情况年度表

软件测试的主要工作集中在测试用例（测试数据）的生成^[216~230]，约简^[231]和优化^[232~235]。其中，利用进化算法进行测试用例生成是目前研究工作的热点，例如：毛澄映等^[216]在讨论基于搜索的测试数据生成基本框架的基础上，以分支覆盖作为测试覆盖准则，给出了基于粒子群优化的测试数据生成算法，并通过分析分支谓词的结构特征提出了一种新的适应函数构造形式；史娇娇等^[217]针对粒子群算法易陷入局部最优解及搜索精度低的问题，提出一种约简的自适应粒子群优化算法应用于测试数据自动生成；姚香娟等^[218]利用语句占有关系，把含有标记变量的测试数据问题转化为语句覆盖问题，然后再利用遗传算法进行求解；张岩等^[219]提出一种基于搜索空间自动缩减的路径覆盖测试数据进化生成方法，使种群在不断缩小的空间里寻找测试数据，以提高测试数据生成的效率；巩敦卫等^[220]采用遗传算法生成回归测试数据以覆盖目标路径时，利用已有测试数据，提出一种新的回归测试数据进化生成方法；侯可佳等^[228]建立了基于服务接口语义契约模型（Interface Semantic Contract，ISC）并探讨了基于 ISC 的测试数据生成技术，给出了分区生成算法以及测试数据生成的模拟退火算法。测试用例约简可有效降低测试用例集规模，顾庆等^[231]针对选择性回归测试中测试需求集部分覆盖要求，提出采用启发式贪婪搜索算法解决多目标部分覆盖测试用例集约简问题。测试用例优化问题是回归测试研究中的一个热点，例如，聂建平等^[234]提出了一种基于 I/O 的 ART 方法，其发现失效的效率较之前方法有了极大的提高，同时可以一次发现多个失效。

国内学者还研究了 SBSE 在 Web 服务的组合优化和网络服务方面的应用^[236~241]，例如：温涛等^[237]提出的改进离散粒子群算法用于解决 Web 服务组合优化问题；谢晓芹等^[239]针对在开放、动态环境下现有的服务发现研究中存在的搜索效率不高、负载不均衡和语义欠缺等问题，提出了一种基于推荐网络和蚁群算法的服务发现方法；黄发良等^[240]将网络社区发现问题形式化为多目标优化问题，提出了一种基于多目标粒子群优化的网络社区发现算法。

此外，基于搜索的方法也在其他方面得到了应用^[242~244]，梁亚澜等^[242]针对已有工作未能系统探索遗传算法生成覆盖表的性能，对遗传算法的配置参数进行了更为深入的探索；严秋玲等^[244]比较了列存储系统中查询优化与行存储系统的不同，在此基础上提出适合于列存储的启发式查询优化机制。

4 国内外研究进展比较

基于搜索的软件工程概念最先由英国 Harman 教授等几位学者提出，并且在英国 EPSRC 的资助下，率先于英国和欧洲召开关于 SBSE 领域研究的研讨会。

图 15 统计了目前已经收录 SBSE 领域文章的作者国家分布，可以看出英国的作者占到四分之一。中国的软件工程学科起步较晚，2012 年才被正式认定为国家一级学科，但中国的学者一直跟踪国际软件工程领域的研究热点，在国际会议和国际期刊发表的文献中作者数量已占到 5% 左右。

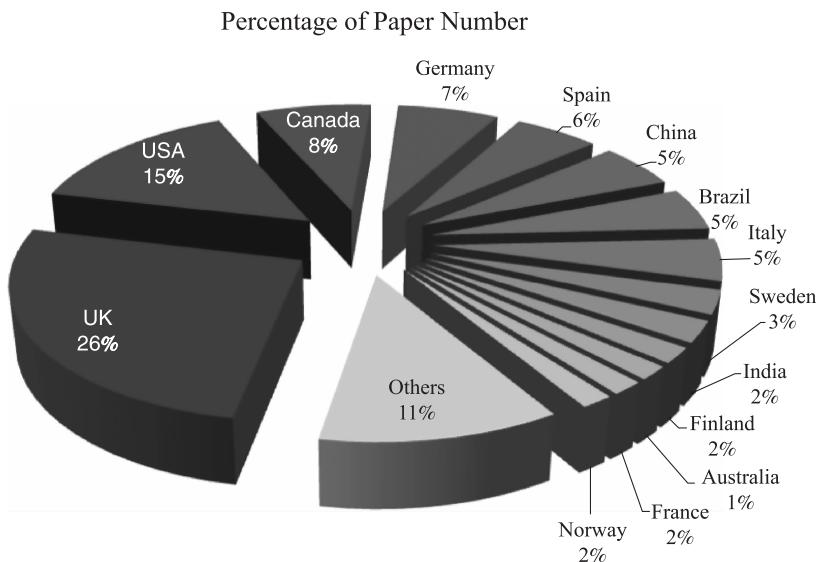


图 15 SBSE 研究人员的国家分布

此外，SBSE 是智能计算与软件工程的交叉，而国内从事智能计算、演化计算、系统优化的人员主要集中在控制科学与工程学科，从事软件工程的研究人员则集中在计算机科学与技术学科，所以目前国内从事 SBSE 领域的研究人员部分是具有计算机学科背景的，部分是具有控制学科背景的。SBSE 真正把两个学科的人交叉凝聚到了一起。

从研究内容上看，由于国内学者密切关注国际上的新热点，同步开展相关研究，在软件测试的自动化和智能化、测试数据自动生成、组合测试、程序自动修复、基于 GPGPU 的并行多目标演化算法等领域有较深入的研究，部分研究成果已经处于世界领先水平。

5 发展趋势与展望

2012 年 Harman 等人在 ACM Computing Surveys 上发表了 SBSE 综述文章 “Search

Based Software Engineering: Trends, Techniques and Applications”^[245]，分析报告了 SBSE 的成果、应用和发展趋势。报告从 SBSE 本质提出四个基础研究方向：

(1) 普适性和可应用性 (Generality and Applicability)

基于搜索的软件工程研究目前主要集中在软件测试领域，但大部分软件工程过程中的行为都可以运用基于搜索的优化方法进行优化。为了达到这一目的，需要解决两方面问题。1) 解决方案表达方式 (Solution Representation)：解决方案的表达方式在最优化问题中扮演一个重要的角色，它部分决定了解决方案的质量，并且帮助软件工程师更好地理解分析结果。不同于其他工程学科，软件工程领域内的问题往往具有较好的可描述性。规范化描述软件工程中更多问题的解决方案，能够使得软件工程更多地受益于基于搜索的优化算法。2) 适应度函数的定义 (Fitness Function Definition)：适应度函数作为评估解决方案的方法，引导了对较优解决方案的搜索。因此，适应度函数是 SBSE 的核心。在软件工程的领域中，评估与度量作为一个重要的研究领域已引起了大量研究人员的研究兴趣，并产生了大量的研究成果。因此，如何对已有的评估方法进行改造，使之转化为可测量的参数，是基于搜索的优化方法在软件工程领域内更广泛应用的重要挑战。

(2) 可扩展性 (Scalability)

软件工程师所面临的一个重要问题就是解决方案有可扩展性，导致许多表现优异的实验室级方法在工业界无法被广泛应用。但幸运的是，基于搜索的优化方法具有天然的并行性。在如今软件高度并行化的趋势、软件工程可扩展性的需要以及基于搜索软件工程方法天然并行性的共同作用下，针对并行的基于搜索的软件工程研究必将得到显著的发展。

(3) 鲁棒性 (Robustness)

在软件工程的一些领域中，问题解决方案的鲁棒性可能与解的优劣同样重要。具体体现为，在搜索空间中找到一个适应度较为稳定的区域要优于找到一个被较差适应度的区域包围的较优解。目前，基于搜索的软件工程的研究关注于构造适应度值更高的解。但是，当目标发生一些微小的变化时，解的适应度可能会显著下降。因此，为了应对这种变化的需要，未来的研究应更多关注于解的鲁棒性。

(4) 反馈与洞察力 (Feedback and Insight)

不同于人为设计，自动化的搜索方法能够避免一些人类直觉可能导致的偏差，这使得基于搜索的方法擅于构造一些意想不到的解，有助于洞察问题解决方案的核心。这一核心优势目前已在工业设计中得到了一定的证实。可以预期，将人为决策与自动化搜索串联结合的迭代过程会产生更好的解决方案。这必将成为未来研究的重要考量。

由于基于搜索的软件测试是 SBSE 研究领域中所占份额最大的部分，下面分 6 个方面给出基于搜索的软件测试的未来研究方向。1) 执行环境的处理，通常基于搜索的测试数据生成技术，缺少对操作系统、文件系统、网络访问及数据库的处理，即软件执行环境的处理；但是执行环境的不同，往往会影响测试的结果，因此有必要对执行环境进行相应的处理。2) 提高改善可测试性的方法，由于标志变量只有两种取值情况，对搜索过程起不到向导的作用，因此通过可测试性转换解决该问题；针对该问题还需要提出更多的

可测试转换方法。3) 通过可测试性转换的自动 Oracles 问题，通过转换生成另一个版本的程序可能不同于原始程序，这就可能引起错误的结果；因此，如何设计适应度函数以区分一个程序的不同版本是值得研究的问题。4) 搜索代码迁徙和重建是否成功，对于不同两个版本的程序，根据它们的行为要采取不同的适应值函数作为向导，因此搜索两个不同元素之间的不同是需要的，其中包括代码的迁徙步骤是否执行正确，及重建是否维持原系统的行为等。5) 最小化人工 Oracle 花费，人工花费往往在测试中占据不小的比例，但是对这方面的关注仍很少；其中一个好的解决方案就是，通过降低生成的测试用例的缺陷检测能力，以降低人工的花费，这便需要二者之间的一个平衡。6) 多个测试目标，在基于搜索的方法中，一次优化多个适应值函数的方法往往应用很少；该方法可以生成同时满足最大覆盖率和最小运行时间的测试数据，对于那些潜在的多个优化目标的问题，都有待于多目标优化算法的应用。

通过对国内外研究的综述与比较，基于搜索的软件工程将呈现以下几方面的趋势：

(1) 大数据环境下的基于搜索的软件工程

伴随着计算机软硬件技术、网络技术、移动通信技术、信息处理技术等蓬勃发展，物联网、云计算、大数据等新技术被业界提出，从不同层面拓展了软件等外延和内涵，也对软件开发提出更多的挑战。2012 年 Harman 等就提出基于云工程也是基于搜索的软件工程^[246]，并通过分析云计算的特点总结出基于搜索的云计算工程所面临的 5 个挑战。在 2014 年的国际基于搜索软件工程研讨会（Symposium on Search- Based Software Engineering, SSBSE 2014）[⊖]上，提出了 SBSE 在大数据领域广泛使用的 Hadoop 系统应用的专题征文。可以看出大数据环境下的基于搜索的软件工程将成为未来的发展趋势之一。

(2) 基于搜索的动态自适应软件工程

SBSE 在软件开发生命周期的每个阶段都有应用，基本涵盖了软件工程的方方面面，但同时也可以看出，现有 SBSE 的应用主要集中在软件生命周期中具体的一个阶段中的个案，并没有面向整个生命周期体系。智能计算及优化在其他工程领域已经有很多系统层次的解决方案，但在软件工程领域还没有，因此我们提出软件工程自动化是 SBSE 的最终目标，基于搜索的动态自适应软件工程将成为 SBSE 未来发展的重要方向之一。

(3) 知识自动化是软件工程自动化的核心

基于搜索的软件工程的最终目标是实现软件工程的自动化，而软件工程自动化的核心就是知识自动化。知识自动化可以分为两个方面，一方面是已知或已约定的知识自动化，另一方面是未知或无法规定的模式的表示及处理，需要融入机器学习和人机交互等方法和技术，间接地改变行为模式，从以“知你为何”为基础实现自动化，转化到以“望你为何”为依据争取智能化，促使希望的测试结果或者目标得以实现。

当前，正值科技革命的重要变革时期，网络空间正以巨大的冲击力影响着我们生活和工作的各个方面。互联网、物联网、云计算、大数据等理念和技术的到来，预示并已经开拓了人类向人工世界进军，深度开发数据和智力资源，深化农业和工业革命

⊖ <http://ssbse.org/2014/>.

的时代使命。在这一历史进程中，以数据驱动的自动化知识将是关键的核心支撑科学和技术。

知识自动化绝对不是知识本身的自动产生，但可以诱发知识的传播、获取、分析、影响、产生等方面的重要变革。知识自动化必将在今后的软件工程中起到关键作用。我们必须从面向物理世界的工业自动化，走向面向数据的知识自动化。

6 结束语

基于搜索的软件工程最终目标是软件工程的自动化和智能化。虽然国内外在该领域的研究已经形成一定规模，但从目前的成果看，离目标还有很大距离，还需要开展更广泛的研究，并进一步结合工业界的实际问题，推动发展和提高水平。

基于搜索的软件工程是智能计算与传统软件工程的交叉与结合。一方面，智能计算、机器学习等技术等发展不断为基于搜索的软件工程领域研究提供新的技术支撑；另一方面，软件开发技术的发展也将影响软件工程的发展，势必为该领域的研究提出更多、更新的挑战，提供更广阔的发展空间。

参考文献

- [1] W Miller, D L. Spooner, Automatic Generation of Floating-Point Test Data [J]. IEEE Transactions on Software Engineering, 1926, 2(3) : 223-226.
- [2] S Xanthakis, C Ellis, C Skourlas, A Le Gall, S Katsikas, K Karapoulios. Application of Genetic Algorithms to Software Testing [C]. Proceedings of the 5th International Conference on Software Engineering and Applications, 1992 : 625-636.
- [3] M Harman, B Jones. Search-based software engineering Information and Software Technology, 2001, 43 (14) : 833-839.
- [4] M Harman. The current state and future of search based software engineering, in Future of Software Engineering 2007 (FOSE 2007) [J]. IEEE Computer Society, 2007 : 342-357.
- [5] M Harman, J Clark. Metrics are fitness functions too, in International Software Metrics Symposium (METRICS 2004) [J]. IEEE Computer Society, 2004 : 58-69.
- [6] Harman M. The current state and future of search based software engineering [C]. 2007 Future of Software Engineering. IEEE Computer Society, 2007 : 342-357.
- [7] E Alba, J F Chicano. Observations in using parallel and sequential evolutionary algorithms for automatic software testing [C]. Computers and Operations Research (COR) focused issue on Search Based Software Engineering.
- [8] J J Dolado. A validation of the component-based method for software size estimation [J]. IEEE Transactions on Software Engineering, 2000, 26(10) : 1006-1021.
- [9] J J Dolado. On the problem of the software cost function [J]. Information and Software Technology, 2001, 43

- (1) : 61-72.
- [10] S Wappler J Wegener. Evolutionary unit testing of object-oriented software using strongly-typed genetic programming[C]. In GECCO 2006: Proceedings of the 8th annual conference on Genetic and evolutionary computation, Seattle, USA, 8-12 July 2006. ACM Press, 2006, 2: 1925-1932.
 - [11] M Harman, R Hierons, M Proctor. A new representation and crossover operator for search-based optimization of software modularization[C]. In GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, USA, 9-13 July 2002. Morgan Kaufmann Publishers, 2002: 1351-1358.
 - [12] B S Mitchell, S Mancoridis. Using heuristic search techniques to extract design abstractions from source code [C]. In GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, USA, 9-13 July 2002. Morgan Kaufmann Publishers, 2002: 1375-1382.
 - [13] S Bouktif, H Sahraoui, G Antoniol. Simulated annealing for improving software quality prediction[C]. In GECCO 2006: Proceedings of the 8th annual conference on Genetic and evolutionary computation, Seattle, USA, 8-12 July 2006. ACM Press, 2006, 2: 1893-1900.
 - [14] M Harman, K Steinhofel, A Skaliotis. Search based approaches to component selection and prioritization for the next release problem [C]. In 22nd International Conference on Software Maintenance (ICSM 06), Philadelphia, USA, Sept. 2006.
 - [15] N Tracey, J Clark, K Mander. Automated program flaw finding using simulated annealing [C]. In International Symposium on Software Testing and Analysis(ISSTA 98). 1998: 73-81.
 - [16] Li H, Lam C P. Software Test Data Generation using Ant Colony Optimization[C]. International Conference on Computational Intelligence. 2004: 1-4.
 - [17] Li H, Lam C P. An ant colony optimization approach to test sequence generation for state based software testing[C]. Quality Software, 2005. (QSIC 2005). Fifth International Conference on. IEEE, 2005: 255-262.
 - [18] Ayari K, Bouktif S, Antoniol G. Automatic mutation test input data generation via ant colony [C]. Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM, 2007: 1074-1081.
 - [19] Windisch A, Wappler S, Wegener J Applying particle swarm optimization to software testing [C]. Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM, 2007: 1121-1128.
 - [20] Li A, Zhang Y. Automatic Generation Method of Test Data for Software Structure Based on PSO [J]. Computer Engineering, 2008, 6: 036.
 - [21] Nie P, Geng J, Qin Z. Self-adaptive inertia weight PSO test case generation algorithm considering prematurity restraining[J]. International Journal of Digital Content Technology and its Applications, 2011, 5(9) : 125-133.
 - [22] John Koza, James Rice. Genetic Generation of Both the Weights and Architecture for a Neural Network[C]. International Joint Conference On Neural Networks, 1991.
 - [23] Jia X, Tang C, Zuo J, Chen A, Duan Lei, Wang R. Mining Frequent Function Set Based on Gene Expression Programming[J]. Chinese Journal of Computers, 2005, 28(8).
 - [24] Xuan S, Liu Y. GEP Evolution Algorithm Based on Control of Mixed Diversity Degree [J]. Pattern Recognition and Artificial Intelligence, 2012, 25(2).
 - [25] Li L, Qu L. Fault Detection Based on Genetic Programming and Support Vector Machines[J]. Journal of

- XIAN Jiaotong University, 2004, 38(3).
- [26] Phil McMinn. Search-based Software Test Data Generation: A survey [J]. *Software Testing, Verification and Reliability*, 2004, 14(2): 105-156.
 - [27] Shaukat Ali, Lionel C Briand, Hadi Hemmati, Rajwinder KaurPanesar-Walawege. A systematic review of the application and empirical investigation of search-based test-case generation [J]. *IEEE Transactions on Software Engineering*, 2010, 36(6): 742-762.
 - [28] Wasif Afzal, Richard Torkar, Robert Feldt. A systematic review of search-based testing for non-functional system properties [J]. *Information and Software Technology*, 2009.
 - [29] Phil McMinn. Search-based software testing: Past, present and future [J]. *Proceedings of the 4th International Workshop on Search-Based Software Testing*, Berlin, Germany, 21-21 March 2011. 2011: 153-163.
 - [30] Webb Miller, David L Spooner. Automatic generation of floating-point test data [J]. *IEEE Transactions on Software Engineering*, 1976, 2(3): 223-226.
 - [31] Bogdan Korel. Automated software test data generation [J]. *IEEE Transactions on Software Engineering*, 1990, 16(8): 870-879.
 - [32] Joachim Wegener, Oliver B. Evaluation of different fitness functions for the evolutionary testing of an autonomous parking system [C]. *Proceedings of the 2004 Conference on Genetic and Evolutionary Computation*, Seattle, USA, 26-30 June 2004. Springer Berlin / Heidelberg, 2004, 3103: 1400-1412.
 - [33] Oliver B, Joachim Wegener. Evolutionary functional testing [J]. *Computers; Operations Research*, 2008, 35(10): 3144-3160.
 - [34] Joachim Wegener, Harmen-Hinrich Sthamer, Bryan F Jones, David E Eyres. Testing real-time systems using genetic algorithms [J]. *Software Quality Journal*, 1997, 6(2): 127-135.
 - [35] Joachim Wegener, Matthias Grochtmann. Verifying timing constraints of real-time systems by means of evolutionary testing [J]. *Real-Time Systems*, 1998, 15(3): 275-298.
 - [36] Joachim Wegener, Frank Mueller. A comparison of static analysis and evolutionary testing for the verification of timing constraints [J]. *Real-Time Systems*, 2001, 21(3): 241-268.
 - [37] Nigel Tracey, John A Clark, Keith Mander. The way forward for unifying dynamic test-case generation: the optimisation-based approach [J]. *Proceedings of the IFIP International Workshop on Dependable Computing and Its Applications*, Johannesburg, South Africa, 12-14 January 1998. 1998: 169-180.
 - [38] Nigel Tracey. A Search-based Automated Test-Data Generation Framework for Safety-Critical Software [D]. University of York, UK, 2000.
 - [39] Hans-Gerhard Groszlig. A prediction system for evolutionary testability applied to dynamic execution time analysis [J]. *Information and Software Technology*, 2001, 43(14): 855-862.
 - [40] Eileen Dillon. Hybrid approach for the automatic determination of worstcase execution time for embedded systems written in c [D]. Institute of Technology, Carlow, 2005.
 - [41] Vu Le Hanh, Kamel Akif, Yves Le Traon, Jean-Marc J. Selecting an efficient object-oriented integration testing strategy: An experimental comparison of actual strategies [C]. *Proceedings of the 15th European Conference on Object-Oriented Programming*, Budapest, Hungary, 18-22 June 2001. Springer, 2001, 2072/2001: 381-401.
 - [42] Lionel C Briand, Jie Feng, Yvan Labiche. Using genetic algorithms and coupling measures to devise optimal integration test orders [C]. *Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering*, Ischia, Italy, 15-19 July 2002. 2002: 43-50.
 - [43] Lionel C Briand, Jie Feng, Yvan Labiche. Experimenting with genetic algorithms and coupling measures to

- devise optimal integration test orders [R]. Technical report, 2003.
- [44] Rafael da Veiga Cabral, Aurora Trinidad Ramirez Pozo, Silvia ReginaVergilio. A pareto ant colony algorithm applied to the class integrationand test order problem [C]. Proceedings of the 22nd IFIP International Conference on Testing Software and Systems, Natal, Brazil, 8-12 November 2010. 2010, 6435 : 16-29.
- [45] Thelma Elita Colanzi, Wesley Klewerton Guez Assuncao, Silvia ReginaVergilio, Aurora Trinidad Ramirez Pozo. Integration test of classesand aspects with a multi-evolutionary and coupling-based approach [C]. Proceedings of the 3rd International Symposium on Search Based Software Engineering, Szeged, Hungary, 10-12 September 2011. 2011, 6956.
- [46] Romain Delamare, Nicholas A Kraft. A genetic algorithm for computingclass integration test orders for aspect-oriented systems [C]. Proceedingsof the 5th International Workshop on Search-Based Software Testing, Montreal, Canada, 21-21 April 2012. 2012: 804-813.
- [47] Wesley Klewerton Guez Assuncao Thelma Elita Colanzi, Silvia ReginaVergilio, Aurora Pozo. On the application of the multi-evolutionaryand coupling-based approach with different aspect-class integration testingstrategies [C]. Proceedings of the 5th International Symposium on SearchBased Software Engineering, St. Petersburg, Russia, 24-26 August 2013. 2013, 8084 : 19-33.
- [48] Lionel C Briand, Yvan Labiche, Marwa Shousha. Stress testing real-timesystems with genetic algorithms [C]. Proceedings of the 2005 Conferenceon Genetic and Evolutionary Computation, Washington, D C, USA, 25-29 June 2005. 2005 : 1021-1028.
- [49] Concettina Del Grosso, Giuliano Antoniol, Massimiliano Di Penta, PhilippeGalinier, Ettore Merlo. Improving network applications security: A new heuristic to generate stress testing data[C]. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation, Washington, D C, USA, 25-29 June 2005. 2005 : 1037-1043.
- [50] Lionel C Briand, Yvan Labiche, Marwa Shousha. Using genetic algorithmsfor early schedulability analysis and stress testing in real-time systems[J]. Genetic Programming and Evolvable Machines, 2006, 7(2) : 145-170.
- [51] Vahid Garousi. Traffic-aware Stress Testing of Distributed Real-Time Systemsbased on UML Models using Genetic Algorithms [D]. Departmentof Systems and Computer Engineering, Carleton University, August 2006.
- [52] Vahid Garousi, Lionel C Briand, Yvan Labiche. Traffic-aware stress testing of distributed real-time systems based on uml models using geneticalgorithms[J]. Journal of Systems and Software, 2008 , 81(2) : 161-185.
- [53] Vahid Garousi. A genetic algorithm-based stress test requirements generatortool and its empirical evaluation [J]. IEEE Transactions on SoftwareEngineering, 2010, 36(6) : 778-797.
- [54] Md. Mehedi Masud, Amiya Nayak, Marzia Zaman, Nita Bansal. Strategyfor mutation testing using genetic algorithms [C]. Proceedings of 2005 Canadian Conference on Electrical and Computer Engineering, Saskatoon, Saskatchewan Canada, 1-4 May 2005. 2005 : 1049-1052.
- [55] Yuan Zhan, John A Clark. Search-based mutation testing for Simulinkmodels[C]. Proceedings of the 2005 Conference on Genetic and EvolutionaryComputation, Washington, D C, USA, 25-29 June 2005. 2005 : 1061-1068.
- [56] Pete May, Jon Timmis, Keith Mander. Immune and evolutionaryapproaches to software mutation testing[C]. Proceedings of the 6th InternationalConference on Artificial Immune Systems, Santos, Brazil, 26-29 August 2007. 2007 : 336-347.

- [57] Yue Jia, Mark Harman. Constructing subtle faults using higher ordermutation testing[C]. Proceedings of the 8th International Working Conferenceon Source Code Analysis and Manipulation, (Best Paper Award), Beijing, China , 28-29 September 2008. 2008 : 249-258 .
- [58] Mark Harman, Yue Jia, William B Langdon. A manifesto for higherorder mutation testing[C]. Proceedings of the 5th International Workshopon Mutation Analysis, Paris, France , 6 April 2010. 2010 : 80-89.
- [59] William B Langdon, Mark Harman, Yue Jia. Multi objective higherorder mutation testing with gp[C]. Proceedings of the 11th Annual Conferenceon Genetic and Evolutionary Computation, Montreal, Canada , 8-12 July 2009. 2009 : 1945-1946.
- [60] JJ Dominguez-Jimenez, A Estero-Botaro, A Garcia-Domnguez, IMedina-Bulo. Evolutionary mutation testing [J]. Information and SoftwareTechnology, 2011.
- [61] Yue Jia, Mark Harman. An analysis and survey of the development ofmutation testing[J]. IEEE Transactions on Software Engineering, 2011 , 37(5) : 649-678.
- [62] Elmahdi Omar, Sudipto Ghosh, Darrell Whitley. Constructing subtlehigher order mutants for java and aspectj programs[C]. Proceedings of IEEE24th International Symposium on SoftwareReliability Engineering, Pasadena, CA, USA , 4-7 November 2013. 2013 : 340-349.
- [63] Myra B Cohen, Charles Joseph Colbourn, Alan C H Ling. Augmentingsimulated annealing to build interaction test suites [C]. Proceedings ofthe 14th International Symposium on Software Reliability Engineering, Denver, Colorado, USA , 17-21 November 2003. 2003 : 394-405.
- [64] S Ghazi, Moataz A Ahmed. Pair-wise test coverage using geneticalgorithms[C]. Proceedings of the IEEE Congress on Evolutionary Computation, Canberra, Australia , 8-12 December 2003. 2013 : 1420-1424.
- [65] Renee C Bryce, Charles Joseph Colbourn, Myra B Cohen. A frameworkof greedy methods for constructing interaction test suites[C]. Proceedingsof the 27th International Conference on Software Engineering, St. Louis, MO, USA , 15-21 May 2005. 2005 : 146-155.
- [66] Renee C Bryce, Charles Joseph Colbourn. Constructing interactiontest suites with greedy algorithms[C]. Proceedings of the 20th IEEE/ACMInternational Conference on Automated Software Engineering, Long Beach, CA, USA , 7-11 November 2005. 2005 : 440-443.
- [67] Renee C Bryce, Charles Joseph Colbourn. One-test-at-a-time heuristic search for interaction test suites[C]. Proceedings of the 9th Annual Conferenceon Genetic and Evolutionary Computation, London, England , 7-11 July 2007. 2007 : 1082-1089.
- [68] Renee C Bryce, Charles Joseph Colbourn. The density algorithm forpairwise interaction testing[J]. Software Testing, Verification and Reliability, 2007 , 17(3) : 159-182.
- [69] Myra B Cohen, Matthew B Dwyer, Jiangfan Shi. Interaction testing ofhighly-configurable systems in the presence of constraints [C]. Proceedingsof the 2007 International Symposium on Software Testing and Analysis, London, United Kingdom , 9-12 July 2007. 2007 : 129-139.
- [70] Myra B Cohen, Matthew B Dwyer, Jiangfan Shi. Constructing interactiontest suites for highly-configurable systems in the presence of constraints: A greedy approach[J]. IEEE Transactions on Software Engineering, 2008 , 34(5) : 633-650.
- [71] Shanghai Nie, Hareton Leung, Baowen Xu. Using computationalsearch to generate 2-way covering array [C]. Proceedings of the 1st InternationalSymposium on Search Based Software Engineering, Cumberland Lodge, Windsor, UK , 13-15 May 2009.
- [72] Brady J Garvin, Myra B Cohen, Matthew B Dwyer. An improvedmeta-heuristic search for constrained

- interaction testing [C]. Proceedings of the 1st International Symposium on Search Based Software Engineering, Cumberland Lodge, Windsor, UK, 13-15 May 2009. 2009 : 13-22.
- [73] Xiang Chen, Qing Gu, Jingxian Qi, Daoxu Chen. Applying particleswarm optimization to pairwise testing [C]. Proceedings of the 34th Annual Computer Software and Applications Conference, Seoul, South Korea, 19-23 July 2010. IEEE, 2010 : 107-116.
- [74] Brady J Garvin, Myra B Cohen, Matthew B Dwyer. Evaluating improvements to a meta-heuristic search for constrained interaction testing[J]. Empirical Software Engineering, 2011 , 16(1) : 61-102.
- [75] Liang Yalan, Changhai Nie, Jonathan M Kauffman, Gregory M Kapfhammer, Hareton Leung. Empirically identifying the best genetic algorithmfor covering array generation[C]. Proceedings of the 3rd International Symposiumon Search Based Software Engineering, Szeged, Hungary, 10-12 September 2011. 2011 , 6956.
- [76] Nashat Mansour, Rami Bahsoon. Reduction-based methods and metricsfor selective regression testing[J]. Information and Software Technology, 2002 , 44(7) : 431-443.
- [77] Ghinwa Baradhi, Nashat Mansour. A comparative study of five regressiontesting algorithms[C]. Proceedings of the Australian Software EngineeringConference, Sydney, NSW, Australia, 28September-2 October 1997. 1997 : 174-182.
- [78] Nashat Mansour, Khalid El-Fakih. Simulated annealing and genetic algorithms for optimal regression testing [J]. Journal of Software Maintenance: Research and Practice, 1999 , 11(1) : 19-34.
- [79] Zheng Li, Mark Harman, Robert M Hierons. Search algorithms for regression test case prioritization[J]. IEEE Transactions on Software Engineering, 2007 , 33(4) : 225-237.
- [80] Shin Yoo. Extending the Boundaries in Regression Testing: Complexity, Latency, and Expertise[D]. PhD thesis, King's College London, UK, 2009.
- [81] Shin Yoo, Mark Harman. Regression testing minimisation, selectionand prioritisation: A survey[J]. Journal of Software Testing, Verification andReliability, 2012 , 22(2) : 67-120.
- [82] Liang You, Yansheng Lu. A genetic algorithm for the time-aware regressiontesting reduction problem[C]. Proceedings of the 8th International Conferenceon Natural Computation, Chongqing, China, 29-31 May 2012. 2012 : 596-599.
- [83] Jeffery Shelburg, Marouane Kessentini, Daniel R. Tauritz. Regressiontesting for model transformations: A multi-objective approach [C]. Proceedings of the 5th International Symposium on Search Based SoftwareEngineering, St. Petersburg, Russia, 24-26 August 2013. 2013 , 8084 : 209-223.
- [84] Joachim Wegener, Andre Baresel, Harmen-Hinrich Sthamer. Evolutionarytest environment for automatic structural testing[J]. Information andSoftware Technology Special Issue on Software Engineering using Meta heuristic Innovative Algorithms, 2001 , 43(14) : 841-854.
- [85] Leonardo Bottaci. Instrumenting programs with flag variables for test datasearch by genetic algorithms[C]. Proceedings of the 2002 Conference onGenetic and Evolutionary Computation, New York, USA, 9-13 July 2002. 2002 : 1337-1342.
- [86] Andre Baresel, Harmen-Hinrich Sthamer. Evolutionary testing of flagconditions[C]. Proceedings of the 2003 Conference on Genetic and EvolutionaryComputation, Chicago, Illinois, USA, 12-16 July 2003. 2003 , 2724 : 2442-2454.
- [87] Mark Harman, Lin Hu, Robert M Hierons, Andre Baresel, Harmen-Hinrich Sthamer. Improving evolutionary testing by flag removal[C]. Proceedings of the 2002 Conference on Genetic and Evolutionary Computation, (Best Paper Award) , New York, USA, 9-13July 2002. 2002 : 1359-1366 .

- [88] Andre Baresel, David Binkley, Mark Harman, Bogdan Korel. Evolutionary testing in the presence of loop-assigned flags: A testability transformation approach [C]. Proceedings of the 2004 ACM SIGSOFT International Symposium on Software Testing and Analysis, Boston, Massachusetts, USA, 11-14 July 2004. ACM, 2004: 108-118.
- [89] Mark Harman, Lin Hu, Robert M Hierons, Joachim Wegener, Harmen-Hinrich Sthamer, Andre Baresel, Marc Roper. Testability transformation [J]. IEEE Transactions on Software Engineering, 2004, 30(1): 3-16.
- [90] Robert M Hierons, Mark Harman, Chris Fox. Branch-coveragetestability transformation for unstructured programs[J]. Computer Journal, 2005, 48(4): 421-436.
- [91] AbdulSalam Kalaji, Robert M Hierons, Stephen Swift. A testabilitytransformation approach for state-based programs[C]. Proceedings of the 1st International Symposium on Search Based Software Engineering, Cumberland Lodge, Windsor, UK, 13-15 May 2009. 2009: 85-88.
- [92] Yanchuan Li, Gordon Fraser. Bytecode testability transformation[C]. Proceedings of the 3rd International Symposium on Search Based Software Engineering, Szeged, Hungary, 10-12 September 2011. 2011, 6956: 237-251.
- [93] Konstantinos Adamopoulos, Mark Harman, Robert M Hierons. How to overcome the equivalent mutant problem and achieve tailored selective mutation using co-evolution[C]. Proceedings of the 2004 Conference on Genetic and Evolutionary Computation, Seattle, Washington, USA, 26-30 June 2004. 2004, 3103/2004: 1338-1349.
- [94] Changhai Nie, Hareton Leung. A survey of combinatorial testing[J]. ACM Computing Surveys, 2011, 43(2): 11: 1-29.
- [95] Paul Baker, Mark Harman, Kathleen Steinhofel, Alexandros Skaliotis. Search basedapproaches to component selection and prioritization for the next release problem[C]. Proceedings of the 22nd IEEE International Conferenceon Software Maintenance, Philadelphia, Pennsylvania, 24-27 September 2006. 2006: 176-185.
- [96] Kristen R. Walcott, Mary Lou Soffa, Gregory M Kapfhammer, Robert S Roos. Time-aware test suite prioritization[C]. Proceedings ofthe 2006 International Symposium on Software Testing and Analysis, Portland, Maine, USA, 17-20 July 2006. 2006: 1-12.
- [97] Praveen Ranjan Srivastava, Aditya Vijay, Bhupesh Barukha, Prashant Singh Sengar, Rajat Sharma. An optimized technique for test casegeneration and prioritization using tabu search and data clustering[C]. Proceedings of the 4th Indian International Conference on Artificial Intelligence, Tumkur, India, 16-18 December 2009. 2009: 30-46.
- [98] Camila Loiola Brito Maia, Fabricio Gomes de Freitas, Jerffeson Teixeirade Souza. Applying search-based techniques for requirements-basedtest case prioritization [C]. Proceedings of the Brazilian Workshop on Optimizationin Software Engineering, Salvador, Brazil, 30-30September 2010.
- [99] Paolo Tonella, Angelo Susi, Francis Palma. Using interactive ga forrequirements prioritization [C]. Proceedings of the 2nd International Symposiumon Search Based Software Engineering, Benevento, Italy, 7-9 September 2010. 2010: 57-66.
- [100] Dayvison Lima, Fabricio Gomes de Freitas, Gustavo Augusto Lima de Campos, Jerffeson Teixeira de Souza. A fuzzy approach to requirementsprioritization[C]. Proceedings of the 3rd International Symposium onSearch Based Software Engineering, Szeged, Hungary, 10-12 September 2011. Springer, 6956: 64-69.
- [101] Sangeeta Sabharwal, Ritu Sibal, Chayanika Sharma. A genetic algorithmbased approach for prioritization of

- test case scenarios in statictesting [C]. Proceedings of the 2nd International Conference on Computerand Communication Technology , Allahabad , India , 15-17 September 2011. 2011 : 304-309.
- [102] Camila Loiola Brito Maia , Thiago do Nascimento Ferreira , Fabricio Gomesde Freitas , Jerffeson Teixeira de Souza . An ant colony based algorithmfor test case prioritization with precedence [C]. Proceedings of the 3rd InternationalSymposium on Search Based Software Engineering , Szeged , Hungary , 10-12 September 2011. 2011 , 6956.
- [103] Matheus Henrique Esteves Paixao , Ma Maria Albuquerque Brasil , Thiago Gomes Nepomuceno da Silva , Jerffeson Teixeira de Souza . Applyingthe ant-q algorithm on the prioritization of software requirements with precedence [C]. Proceedings of the 3rd Brazilian Workshop on Search-BasedSoftware Engineering , Natal , RN , Brazil , 23 September 2012.
- [104] Zheng Li , Yi Bian , Ruilian Zhao , Jun Cheng . A fine-grained parallelmulti-objective test case prioritization on gpu [C]. Proceedings of the 5thInternational Symposium on Search Based Software Engineering , St. Petersburg , Russia , 24-26 August 2013. 2013 , 8084 : 111-125.
- [105] Paolo Tonella , Angelo Susi , Francis Palma . Interactive requirementsprioritization using a genetic algorithm [J]. Information and Software Technology , 2013 , 55(1) : 173-187.
- [106] Antonio Mauricio Pitangueira , Rita Suzana P Maciel , Marcio de OliveiraBarros , Aline Santos Andrade . A systematic review of softwarerequirements selection and prioritization using sbse approaches [C]. Proceedings of the 5th International Symposium on Search Based SoftwareEngineering (SSBSE '13) , St. Petersburg , Russia , 24-26 August 2013. 2013 , 8084 : 188-208.
- [107] Phil McMinn . Search-based failure discovery using testability transformationsto generate pseudo-oracles [C]. Proceedings of the 11th Annual Conferenceon Genetic and Evolutionary Computation , Montreal , Canada , 8-12 July 2009. 2009 : 1689-1696.
- [108] Mark Harman , Sung Gon Kim , Kiran Lakhota , Phil McMinn , ShinYoo . Optimizing for the number of tests generated in search based test datageneration with an application to the oracle cost problem [C]. Proceedings ofthe 3rd International Workshop on Search-Based Software Testingin conjunction with ICST 2010 , Paris , France , 6 April 2010. 2010 : 182-191.
- [109] Gordon Fraser , Andrea Arcuri . Whole test suite generation [J]. IEEE Transactions on Software Engineering , 2013 , 39 : 276-291.
- [110] Phil McMinn , Mark Stevenson , Mark Harman . Reducing qualitativehuman oracle costs associated with automatically generated test data [C]. Proceedings of the 1st International Workshop on Software Test OutputValidation , Trento , Italy , 13 July 2010. 2010 : 1-4.
- [111] Reza Matinnejad , Shiva Nejati , Lionel Briand , Thomas Bruckmann , Claude Poull . Search-based automated testing of continuous controllers: Framework , tool support , and case studies [J/OL]. Information and Software Technology , Available online 22 May 2014 , ISSN 0950-5849. <http://dx.doi.org/10.1016/j.infsof.2014.05.007>.
- [112] Yuhua Qi , Xiaoguang Mao , Yan Lei , Ziying Dai , Chengsong Wang . The Strength of Random Search on Automated Program Repair [C]. 36th International Conference on Software Engineering , ICSE 2014 , May 31-June 7 , Hyderabad , India , 2014.
- [113] Qi Yuhua , Mao Xiaoguang , Lei Yan , Wang Chensong . Using automated program repair for evaluating the effectiveness of fault localization techniques [C]. 22nd International Symposium on Software Testing and Analysis , ISSTA 2013 , July 15-20 , pp. 191-201 , Lugano , Switzerland , 2013.

- [114] Qi YuHua, Mao XiaoGuang, Wen YanJun, Dai ZiYing, Gu Bin. More efficient automatic repair of large-scale programs using weak recompilation [J]. *Science China-Information Sciences*, 2012, 55 (12) : 2785-2799.
- [115] Qi Yuhua, Mao Xiaoguang, Lei Yan. Making automatic repair for large-scale programs more efficient using weak recompilation [C]. 28th IEEE International Conference on Software Maintenance, ICSM 2012, 2012/9/23-2012/9/28, pp. 254-263, Riva del Garda, Trento, Italy, 2012.
- [116] Qi Yuhua, Mao Xiaoguang, Lei Yan. Efficient automated program repair through fault-recorded testing prioritization [C]. 29th IEEE International Conference on Software Maintenance, ICSM 2013, September 22-28, pp. 180-189, Eindhoven, Netherlands, 2013.
- [117] Mao Xiaoguang, Lei Yan, Dai Ziying, Qi Yuhua, Wang Chengsong. Slice-based statistical fault localization [J]. *Journal of Systems and Software*, 2014, 89 (1) : 51-62.
- [118] Lei Yan, Mao Xiaoguang, Dai Ziying, Wang Chengsong. Effective statistical fault localization using program slices [C]. 36th IEEE Annual International Computer Software and Applications Conference, COMPSAC 2012, July 16-20, pp. 1-10, Izmir, Turkey, 2012.
- [119] Lei Yan, Mao Xiaoguang, Chen Tsong Yueh. Backward-slice-based statistical fault localization without test oracles [C]. 13th International Conference on Quality Software, QSIC 2013, July 29-30, pp. 212-221, Nanjing, Jiangsu, China, 2013.
- [120] Bagnall A J, Rayward-Smith V J, Whittle I M. The Next Release Problem [J]. *Information & Software Technology*, 2001, 43 (14) : 883-890.
- [121] Zhang Y, Harman M, Mansouri S A. The multi-objective next release problem [C]. International Conference on Genetic and Evolutionary Computation (GECCO2007), New York: ACM, 2007: 1129-1136.
- [122] Saliu M O, Ruhe G. Bi-Objective Release Planning for Evolving Software Systems [C]. 6th European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering, New York : ACM, 2007: 105-114.
- [123] Y Zhang, M Harman, Search based optimization of requirements interaction management [C]. Proceedings of the 2nd International Symposium on Search Based Software Engineering (SSBSE '10), Benevento, Italy. IEEE, 2010: 47-56.
- [124] Zhang Y. Multi-Objective Search-based Requirements Selection and Optimisation [D]. University of London, 2010.
- [125] Tonella P, Susi A, Palma F. Interactive requirements prioritization using a genetic algorithm [J]. *Information and Software Technology*, 2013, 55 (1) : 173-187.
- [126] Tonella P, Susi A, Palma F. Using interactive GA for requirements prioritization [C]. Search Based Software Engineering (SSBSE), 2010 Second International Symposium on. IEEE, 2010: 57-66.
- [127] Kumari A C, Srinivas K, Gupta M P. Software requirements selection using Quantum-inspired Elitist Multi-objective Evolutionary algorithm [C]. Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on. IEEE, 2012: 782-787.
- [128] Jifeng Xuan, He Jiang, Zhilei Ren, Zhongxuan Luo. Solving the Large Scale Next Release Problem with a Backbone Based Multilevel Algorithm. *IEEE Transactions on Software Engineering*, vol. 38, no. 5, Sept. - Oct. 2012, pp. 1195-1212
- [129] Kumari A C, Srinivas K, Gupta M P. Software Requirements Optimization Using Multi-Objective Quantum-Inspired Hybrid Differential Evolution [M]. EVOLVE-A Bridge between Probability, Set Oriented

- Numerics, and Evolutionary Computation II. Springer Berlin Heidelberg, 2013: 107-120.
- [130] Gay G, Menzies T, Jalali O, et al. Finding robust solutions in requirements models [J]. Automated Software Engineering, 2010, 17(1): 87-116.
- [131] Zhang Y, Harman M, Finkelstein A, et al. Comparing the performance of metaheuristics for the analysis of multi-stakeholder tradeoffs in requirements optimisation [J]. Information and Software Technology, 2011, 53(7): 761-773.
- [132] Veerappa V. Clustering Methods for Requirements Selection and Optimisation[D]. UCL(University College London), 2013.
- [133] Pitangueira A M, Maciel R S P, de Oliveira Barros M, et al. A systematic review of software requirements selection and prioritization using SBSE approaches [M]. Search Based Software Engineering. Springer Berlin Heidelberg, 2013: 188-208.
- [134] Ren J: Sensitivity Analysis in Multi-Objective Next Release Problem and Fairness Analysis in Software Requirements Engineering[D]. Master's thesis, DCS/PSE, King's College London, London, 2007.
- [135] A Finkelstein, M Harman, A Mansouri, J Ren, Y Zhang. Fairness analysis in requirements assignments [C]. 16th IEEE International Requirements Engineering Conference, Los Alamitos, California, USA. IEEE Computer Society Press, 2008.
- [136] Y Zhang, A Finkelstein, M Harman. Search based requirements optimisation: Existing work and challenges [C]. International Working Conference on Requirements Engineering: Foundation for Software Quality (REFSQ '08), Montpellier, France. Springer LNCS, 2008.
- [137] Yuanyuan Zhang, Mark Harman, Soo Ling Lim, Empirical evaluation of search based requirements interaction management [J]. Information and Software Technology, ISSN 0950-5849. 2013, 55 (1): 126-152.
- [138] Outi Raiha. A survey on search-based software design [J]. Computer Science Review, 2010, 4 (4): 203-249.
- [139] Christopher L. Simons and Ian C Parmee. A cross-disciplinary technologytransfer for search-based evolutionary computing: from engineering design to software engineering design [J]. Engineering Optimization, 2007, 39(5): 631-648.
- [140] Christopher L. Simons, Ian C Parmee. Single and multi-objectivegenetic operators in object-oriented conceptual software design [C]. Proceedings of the 8th annual Conference on Genetic and Evolutionary Computation, pages 1957-1958, Seattle, Washington, USA, 8-12July 2006.
- [141] Christopher L Simons, Ian C Parmee. User-centered, evolutionary search in conceptual software design [C]. Proceedings of the IEEE Congress on Evolutionary Computation, pages 869-876, Hong Kong, China, 1-6 June 2008.
- [142] Mark O'Keeffe, Mel O Cinneide. A stochastic approach to automateddesign improvement[C]. Proceedings of the 2nd International Conference on Principles, Practice of Programming in Java, pp. 59-62, Kilkenny City, Ireland, 16-18 June 2003.
- [143] Mark O'Keeffe, Mel O Cinneide. Towards automated design improvementthrough combinatorial optimisation [C]. Proceedings of the 26th International Conference on Software Engineering and Workshop on Directionsin Software Engineering Environments, pages 75-82, Edinburgh, UK, 23-28 May 2004.
- [144] Michael Bowman, Lionel C Briand, Yvan Labiche. Solving the class responsibility assignment problem in object-oriented analysis with multiobjectivegenetic algorithms [R]. Technical Report SCE-07-02,

August 2008.

- [145] Mehdi Amoui, Siavash Mirarab, Sepand Ansari, Caro Lucas. A geneticalgorithm approach to design evolution using design pattern transformation [J]. International Journal of Information Technology and IntelligentComputing , 2006, 1(2) : 235-244.
- [146] Outi Raiha. Applying genetic algorithms in software architecture design[D]. Master's thesis , Department of Computer Sciences , University of Tampere , February 2008.
- [147] Outi Raiha. Genetic Synthesis of Software Architecture[D]. PhD thesis , Universityof Tampere , Finland , September 2008.
- [148] Heather J Goldsby, Betty H C Cheng. Automatically generating behavioral models of adaptive systems to address uncertainty[C]. Proceedingsof the 11th International Conference on Model Driven Engineering Languages and Systems , pages 568-583 , Toulouse , France , 28September-3 October 2008.
- [149] Heather J Goldsby, Betty H C Cheng. Avida-mde: a digital evolutionapproach to generating models of adaptive software behavior [C]. Proceedings of the 10th Annual Conference on Genetic and EvolutionaryComputation , pages 1751-1758 , Atlanta , GA , USA , 12-16July 2008.
- [150] Heather J Goldsby, Betty H C Cheng, Philip K. McKinley, David B Knoester, Charles A Ofria. Digital evolution of behavioral models forautonomic systems [C]. Proceedings of the 2008 International Conferenceon Autonomic Computing , pages 87-96 , Chicago , IL , USA , 2-6June 2008.
- [151] Simon M Lucas, T. Jeff Reynolds. Learning deterministic finite automatawith a smart state labeling evolutionary algorithm[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2005 , 27 (7) : 1063-1074.
- [152] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, MariaLuisa Villani. An approach for qos-aware service composition based ongenetic algorithms[C]. Proceedings of the 2005 Conference on Genetic and Evolutionary Computation , pages 1069-1075 , Washington , D C , USA , 25-29 June 2005.
- [153] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, MariaLuisa Villani. Qos-aware replanning of composite web services[C]. Proceedings of 2005 IEEE International Conference on Web Services , pages 121-129 , Orlando , FL , USA , 11-15 July 2005.
- [154] Michael C Jaeger, Gero Muhl. Qos-based selection of services: Theimplementation of a genetic algorithm [C]. Proceedings of Kommunikationin Verteilten Systemen2007 Workshop: Service-Oriented Architectures and Service-Oriented Computing , 2007.
- [155] Taghi M Khoshgoftaar, Yi Liu, Naeem Seliya. A multiobjectivemode-order model for software quality enhancement[J]. IEEE Transactionson Evolutionary Computation , 2004 , 8(6) : 593-608.
- [156] Taghi M Khoshgoftaar, Yi Liu, Naeem Seliya. Module-order modeling using an evolutionary multi-objective optimization approach[C]. Proceedings of the 10th IEEE International Symposium on Software Metrics , pages 159-169 , Chicago , USA , 14-16 September 2004.
- [157] Chengwen Zhang, Sen Su, Junliang Chen. A novel genetic algorithm for qos-aware web services selection [C]. Proceedings of the 2nd International Workshop on Data Engineering Issues in E-Commerce and Services , volume 4055 , pages 224-235 , San Francisco , CA , USA , 26 June 2006.
- [158] Chengwen Zhang, Sen Su, Junliang Chen. Diga: Population diversityhandling genetic algorithm for qos-aware web services selection[J]. ComputerCommunications , 2007 , 30(5) : 1082-1090.
- [159] Sen Su, Chengwen Zhang, Junliang Chen. An improved genetic algorithm for web services selection[C]. Proceedings of the 7th IFIP WG6. 1 International Conference on Distributed Applications and Interoperable

- Systems, Paphos, Cyprus, 6-8 June 2007. 2007, 4531 : 284-295.
- [160] Yue Ma, Chengwen Zhang. Quick convergence of genetic algorithmfor qos-driven web service selection[J]. Computer Networks, 2008 , 52(5) : 1093-1104.
- [161] Lei Cao, Minglu Li, Jian Cao. Cost-driven web service selection usinggenetic algorithm[C]. Proceedings of the 1st International Workshop onInternet and Network Economics, pages 906-915 , Hong Kong, China, 15-17 December 2005.
- [162] K Vijayalakshmi, N Ramaraj, R Amuthakkannan. Improvement ofcomponent selection process using genetic algorithm for component-basedsoftware development [J]. International Journal of Information Systems andChange Management, 2008 , 3(1) : 63-80.
- [163] Lili Yang, Bryan F Jones, Shuang-Hua Yang. Genetic algorithm based software integration with minimum software risk[J]. Information and Software Technology, 2006 , 48(3) : 133-141.
- [164] Nicolas Desnos, Marianne Huchard, Guy Tremblay, Christelle Urtado, Sylvain Vauttier. Search-based many-to-one component substitution[J]. Journal of Software Maintenance and Evolution: Research and Practice, 2008 , 20(5) : 321-344.
- [165] Vittorio Cortellessa, Ivica Crnkovic, Fabrizio Marinelli, PasqualinaPotena. Experimenting the automated selection of cots components basedon cost and system requirements[J]. Journal of Universal Computer Science, 2008 , 14(8) : 1228-1255.
- [166] Michael Kuperberg, Klaus Krogmann, Ralf Reussner. Performance prediction for black-box components using reengineered parametric behavior models[C]. Proceedings of the 11th International Symposium on Component-Based Software Engineering, Karlsruhe, Germany, 14-17 October 2008. 2008 , 5282 : 48-63.
- [167] Robert Feldt. An experiment on using genetic programming to developmultiple diverse software variants [R]. Technical Report 98-13 , Gothenburg, Sweden, September 1998.
- [168] Robert Feldt. Generating multiple diverse software versions with genetic programming[C]. Proceedings of the 24th EUROMICRO Conference, Vasteras, Sweden, 25-27 August 1998. 1998 , 1 : 387-394.
- [169] Robert Feldt. Generating multiple diverse software versions with geneticprogramming-an experimental study [C]. IEE Proceedings-Software, 1998 , 145(6) : 228-236.
- [170] Gerassimos Barlas, Khaled El-Fakih. A ga-based movie-on-demandplatform using multiple distributed servers[J]. Multimedia Tools and Applications, 2008 , 40(3) : 361-383.
- [171] Sylvain Chardigny, Abdelhak Seriai, Dalila Tamzalit, Mourad Oussalah. Quality-driven extraction of a component-based architecture froman object-oriented system [C]. Proceedings of the 12th European Conference on Software Maintenance and Re engineering, pages 269-273 , Athens, Greece, 1-4 April 2008.
- [172] Vibhu Saujanya Sharma, Pankaj Jalote. Deploying software componentsfor performance[C]. Proceedings of the 11th International Symposiumon Component-Based Software Engineering, pages 32-47 , Karlsruhe, Germany, 14-17 October 2008.
- [173] C Ryan. Automatic re-engineering of software using genetic programming[J]. Springer, 1999 , 2.
- [174] K P Williams. Evolutionary algorithms for automatic parallelization[D]. Doctoral dissertation, University of Reading, 1998.
- [175] D Fatiregun, M Harman, R M Hierons. Evolving transformation sequences using genetic algorithms[J]. Source Code Analysis and Manipulation, 2004 , pages 65-74 , Sep. 2004.
- [176] M O'Keeffe, M O'Cinneide. Search-based software maintenance[C]. Conference on Software Maintenance

- and Reengineering(CSMR '06) , pages 249-260 , Mar. 2006.
- [177] S Bouktif, G. Antoniol, E Merlo, M Neteler. A novel approach to optimize clone refactoring activity[C]. Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation, pages 1885-1892, July. 2006.
- [178] O Seng, M Bauer, M Biehl, G. Pache(2005 , June). Search-based improvement of subsystem decompositions [C]. Proceedings of the 2005 conference on Genetic and evolutionary computation . ACM, 2005: 1045-1051.
- [179] A Ghannem, G El Boussaidi, M Kessentini. Model refactoring using interactive genetic algorithm[C]. Search Based Software Engineering(SBSE '2013) , pages 96-110, Springer Berlin Heidelberg.
- [180] R Lämmel. Towards generic refactoring[C]. Proceedings of the 2002 ACM SIGPLAN workshop on Rule-based programming . 2002: 15-28.
- [181] H Li, S Thompson. Comparative study of refactoring haskell and erlang programs[C]. Source Code Analysis and Manipulation, (SCAM '2006) , pages 197-206, Sep. 2006.
- [182] M Harman. Open problems in testability transformation[C]. Software Testing Verification and Validation Workshop, (ICSTW'2008) , pages 196-209 , April 2008.
- [183] M Harman. Refactoring as testability transformation[C]. Software Testing, Verification and Validation Workshops (ICSTW '2011) , pages 414-421 , Mar. 2011.
- [184] M Harman, L Tratt. Pareto optimal search based refactoring at the design level[C]. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO' 2007) , pages 1106-1113, July 2007.
- [185] M K O'Keeffe, M O Cinneide. Getting the most from search-based refactoring[C]. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO' 2007) , pages 1114-1120, July 2007.
- [186] E Koc, N Ersoy, A Andac, Z S Camlidere, I Cereci, H Kilic. An empirical study about search-based refactoring using alternative multiple and population-based search techniques [J]. Computer and Information Sciences II, London: Springer, 2012: 59-66.
- [187] J Tao, N Gold, M Harman, Z Li. Locating dependence structures using search based slicing [J]. Information and Software Technology.
- [188] N Gold, M Harman, Z Li, K Mahdavi. A search based approach to overlapping concept boundaries[C]. 22nd International Conference on Software Maintenance (ICSM 06) , pages 310-319, Philadelphia, Pennsylvania, USA , Sept. 2006.
- [189] Alvarez-Valdes R, Crespo E, Tamarit JM, Villa F. A scatter search algorithm for project scheduling under partially renewable resources[J]. Journal of Heuristics, 2006, 12(1-2): 95-113.
- [190] G. Antoniol, M Di Penta, M Harman. A robust search-based approach to project management in the presence of abandonment, rework, error and uncertainty [C]. 10th International Software Metrics Symposium(METRICS 2004) , Los Alamitos, California, USA , Sept. 2004. IEEE Computer Society Press, 2004: 172-183.
- [191] Di Penta M, HarmanM, Antoniol G. The use of search-based optimization techniques to schedule and staff software projects: an approach and an empirical study[J]. Software: Practice and Experience, 2011, 41 (5): 495-519.
- [192] Jifeng Xuan, He Jiang, Zhilei Ren, Weiqin Zou. Developer Prioritization in Bug Repositories [C].

- Proceedings of 34th International Conference on Software Engineering (ICSE 2012), Zurich, Switzerland. June 2-9, 2012. 2012: 25-35.
- [193] Di Penta M, Harman M, Antoniol G, Qureshi F. The effect of communication overhead on software maintenance project staffing: a search-based approach [C]. In Software Maintenance, 2007. ICSM 2007. IEEE, 2007: 315-324.
- [194] Stylianou C, Andreou A S. A multi-objective genetic algorithm for intelligent software project scheduling and team staffing [J]. Intelligent Decision Technologies, 2013, 7(1): 59-80.
- [195] Chicano F, Luna F, Nebro A J, Alba E. Using multi-objective metaheuristics to solve the software project scheduling problem [C]. Proceedings of the 13th annual conference on Genetic and evolutionary computation . ACM, 2012, 1915-1922.
- [196] Chen W N, Zhang J. Ant colony optimization for software project scheduling and staffing with an event-based scheduler [J]. Software Engineering, IEEE Transactions on, 2013, 39(1): 1-17.
- [197] Burgess C J, Lefley M. Can genetic programming improve software effort estimation? A comparative evaluation [J]. SERIES ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 2005, 16, 95.
- [198] Bardsiri V K, Jawawi D N A, Hashim S ZM, Khatibi E. A PSO-based model to increase the accuracy of software development effort estimation [J]. Software Quality Journal, 2013, 21(3): 501-526.
- [199] GAntoniol, M D Penta, M Harman. Search-based techniques applied to optimization of project planning for a massive maintenance project [C]. 21st IEEE International Conference on Software Maintenance, Los Alamitos, California, USA, 2005. IEEE Computer Society Press, 2005: 240-249.
- [200] Dolado J J. On the problem of the software cost function [J]. Information and Software Technology, 2001, 43(1): 61-72.
- [201] Azar D. A Genetic Algorithm for Improving Accuracy of Software Quality Predictive Models: A Search-based Software Engineering Approach [J]. International Journal of Computational Intelligence and Applications, 2010, 9(02): 125-136.
- [202] J Aguilar-Ruiz, I Ramos, J C Riquelme, M Toro. An evolutionary approach to estimating software development projects [J]. Information and Software Technology, 2001, 43(14): 875-882.
- [203] C Kirsopp, M Shepperd, J Hart. Search heuristics, case-based reasoning and software project effort prediction [C]. GECCO 2002: Proceedings of the Genetic and Evolutionary Computation Conference, San Francisco, CA 94104, USA, 9-13 July 2002. Morgan Kaufmann Publishers, 2002: 1367-1374.
- [204] J Wegener A Baresel, H Stamer. Evolutionary Test Environment for Automatic Structural Testing [J]. Information and Software Technology, 2001, 43(14): 841-854.
- [205] Windisch A Wappler S, Wegener J. Applying particle swarm optimization to software testing [C]. Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation. 2007.
- [206] Harman M, Hassoun Y, Lakhotia K, McMinn P, Wegener J. The impact of input domain reduction on search-based test data generation [C]. Proceedings of the 6th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering. New York: ACM Press, 2007: 1121-1128.
- [207] Buhler O, Wegener J. Evolutionary functional testing [J]. Comput. Oper. Res. 35, 10, 3144-3160. 2008.
- [208] C Cadar, P Godefroid, S Khurshid, C S Pasareanu, K Sen, N Tillmann, W Visser. Symbolic Execution for Software Testing in Practice: Preliminary Assessment [C]. Proceedings of the 33rd International

Conference on Software Engineering, pp. 1066- 1071, Hawaii, USA, 21-28 May 2011.

- [209] Xie T, Tillmann N, Dehalleux P, SchulteW. Fitness-Guided path exploration in dynamic symbolic execution [R]. Tech. rep. MSR-TR-2008-123, Microsoft Research. September. 2008.
- [210] K Lakhotia, N Tillmann, M Harman, J de Halleux. FloPSy Search-Based Floating Point Constraint Solving for Symbolic Execution [C]. Proceedings of the 22nd IFIP International Conference on Testing Software and Systems, Natal, Brazil, November 2010. Springer, 2010, 6435 (of LNCS) : 142- 157.
- [211] S Yoo, R Nilsson M Harman. Faster Fault Finding at Google using Multi Objective Regression Test Optimisation [C]. Proceedings of the 8th European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Szeged, Hungary, 5-9 September 2011.
- [212] Yoo S, Harman M, Ur S. Measuring and improving latency to avoid test suite wear out [C]. Proceedings of the IEEE International Conference on Software Testing, Verification, and Validation Workshops. IEEE, 2009: 101-110.
- [213] P Baker, M HarmanK Steinhofel, A Skaliotis. Search Based Approaches to Component Selection and Prioritization for the Next Release Problem [C]. Proceedings of the 22nd International Conference on Software Maintenance, pp. 176-185, Philadelphia, Pennsylvania, USA, 24-27 September 2006.
- [214] Y Zhang, E Alba, J J Durillo, S Eldh, M Harman. Today/Future Importance Analysis [C]. Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation, pp. 1357-1364, Portland, Oregon, USA, 7-11 July 2010.
- [215] Kiran Lakhotia, Mark Harman, Hamilton Gross. AUSTIN: An open source tool for search based software testing of C programs [J]. Information and Software Technology, 2013, 55(1) : 112-125.
- [216] 毛澄映, 喻新欣, 薛云志. 基于粒子群优化的测试数据生成及其实证分析 [J]. 计算机研究与发展, 2014, 04: 824-837.
- [217] 史娇娇, 姜淑娟, 韩寒, 王令赛. 自适应粒子群优化算法及其在测试数据生成中的应用研究 [J]. 电子学报, 2013, 08: 1555-1559.
- [218] 姚香娟, 巩敦卫. 基于目标语句占优关系的软件可测试性转化 [J]. 电子学报, 2013, 12: 2523-2528.
- [219] 张岩, 巩敦卫. 基于搜索空间自动缩减的路径覆盖测试数据进化生成 [J]. 电子学报, 2012, 05: 1011-1016.
- [220] 巩敦卫, 任丽娜. 回归测试数据进化生成 [J]. 计算机学报, 2014, 03: 489-499.
- [221] 巩敦卫, 张岩. 一种新的多路径覆盖测试数据进化生成方法 [J]. 电子学报, 2010, 06: 1299-1304.
- [222] 舒挺, 刘良桂, 徐伟强, 李文书. 自适应 EFSM 可执行测试序列生成 [J]. 计算机研究与发展, 2012, 06: 1211-1219.
- [223] 刘新忠, 徐高潮, 胡亮, 付晓东, 董玉双. 一种基于约束的变异测试数据生成方法 [J]. 计算机研究与发展, 2011, 04: 617-626.
- [224] 侯可佳, 白晓颖, 陆皓, 李树芳, 周立柱. 基于接口语义契约的 Web 服务测试数据生成 [J]. 软件学报, 2013, 09: 2020-2041.
- [225] 谢晓园, 徐宝文, 史亮, 聂长海. 面向路径覆盖的演化测试用例生成技术 (英文) [J]. 软件学报, 2009, 12: 3117-3136.
- [226] 王建民, 蔡媛. 基于维持种群多样性的测试数据生成算法的研究 [J]. 计算机研究与发展, 2012, 05: 1039-1048.
- [227] 张岩, 巩敦卫. 基于稀有数据捕捉的路径覆盖测试数据进化生成方法 [J]. 计算机学报, 2013, 12: 2429-2440.

- [228] 田甜, 巩敦卫. 消息传递并行程序路径覆盖测试数据生成问题的模型及其进化求解方法[J]. 计算机学报, 2013, 11: 2212-2223.
- [229] 查日军, 张德平, 聂长海, 徐宝文. 组合测试数据生成的交叉熵与粒子群算法及比较[J]. 计算机学报, 2010, 10: 1896-1908.
- [230] 董国伟, 聂长海, 徐宝文. 基于程序路径分析的有效蜕变测试[J]. 计算机学报, 2009, 05: 1002-1013
- [231] 顾庆, 唐宝, 陈道蓄. 一种面向测试需求部分覆盖的测试用例集约简技术[J]. 计算机学报, 2011, 05: 879-888.
- [232] 陈翔, 陈继红, 鞠小林, 顾庆. 回归测试中的测试用例优先排序技术述评[J]. 软件学报, 2013, 08: 1695-1712.
- [233] 张智轶, 陈振宇, 徐宝文, 杨瑞. 测试用例演化研究进展[J]. 软件学报, 2013, 04: 663-674.
- [234] 聂剑平, 曹旭, 钱越英, 陈昱松. 基于 I/O 关系的适应性随机测试[J]. 计算机研究与发展, 2010, S1: 56-63.
- [235] 夏亚梅, 程渤, 陈俊亮, 孟祥武, 刘栋. 基于改进蚁群算法的服务组合优化[J]. 计算机学报, 2012, 02: 2270-2281.
- [236] 王尚广, 孙其博, 杨放春. 基于全局 QoS 约束分解的 Web 服务动态选择[J]. 软件学报, 2011, 07: 1426-1439.
- [237] 温涛, 盛国军, 郭权, 李迎秋. 基于改进粒子群算法的 Web 服务组合[J]. 计算机学报, 2013, 05: 1031-1046.
- [238] 邓亮, 赵进, 王新. 基于遗传算法的网络编码优化[J]. 软件学报, 2009, 08: 2269-2279.
- [239] 谢晓芹, 宋超臣, 张志强. 一种基于推荐网络和蚁群算法的服务发现方法[J]. 计算机学报, 2010, 11: 2093-2103.
- [240] 黄发良, 张师超, 朱晓峰. 基于多目标优化的网络社区发现方法[J]. 软件学报, 2013, 09: 2062-2077.
- [241] 曾明霏, 余顺争. P2P 网络服务器部署方案及其启发式优化算法[J]. 软件学报, 2013, 09: 2226-2237.
- [242] 梁亚澜, 聂长海. 覆盖表生成的遗传算法配置参数优化[J]. 计算机学报, 2012, 07: 1522-1538.
- [243] 聂长海, 蒋静. 覆盖表生成的可配置贪心算法优化[J]. 软件学报, 2013, 07: 1469-1483.
- [244] 严秋玲, 孙莉, 王梅, 乐嘉锦, 刘国华. 列存储数据仓库中启发式查询优化机制[J]. 计算机学报, 2011, 10: 2018-2026.
- [245] Mark Harman, Afshin Mansouri, Yuanyuan Zhang. Search Based Software Engineering: Trends, Techniques and Applications[J]. ACM Computing Surveys, 2012, 45(1): Article 11.
- [246] Mark Harman, Kiran Lakhotia, Jeremy Singer, David R White, Shin Yoo, Cloud engineering is Search Based Software Engineering too[J]. Journal of Systems and Software, 2013, 86(9): 2225-2241.

附录：基于搜索的软件工程相关资源

1. 基于软件工程的文献库：<http://www.sebase.orgsbsepublications/>
2. 基于搜索的软件工程领域的相关综述文章（Survey）

- 测试数据生成

Phil McMinn. Search- based software test data generation: a survey [J]. *Software Testing, Verification and Reliability*, 2004, 14 (2): 105-156.

Shaukat Ali, Lionel Briand, Hadi Hemmati and Rajwinder Panesar- Walawege. A Systematic Review of the Application and Empirical Investigation of Search- Based Test- Case Generation [J]. *IEEE Transactions on Software Engineering*, 2010, 36 (6): 742-762.

Wasif Afzal, Richard Torkar and Robert Feldt. A Systematic Review of Search- based Testing for Non- Functional System Properties [J]. *Information and Software Technology*, 2009, 51 (6): 957-976.

- 软件设计

Outi Räihä. A Survey on Search- Based Software Design [J]. *Computer Science Review*, 2010, 4 (4): 203-249.

- 发展趋势

Mark Harman. The Current State and Future of Search Based Software Engineering [C]. *ICSE Future of Software Engineering*, 2007: 342-357.

Mark Harman, Afshin Mansouri and Yuanyuan Zhang. Search Based Software Engineering: Trends, Techniques and Applications [J]. *ACM Computing Surveys*, 2012, 45 (1): Article 11.

3. 基于搜索的软件工程相关工具

- Austin 针对 C 程序的，基于搜索的软件测试数据生成工具。

<http://code.google.com/p/austin-sbst/>

- Cocotest 利用搜索的方法对连续控制器进行自动化的 MIL (Model-In-the-Loop) 测试的工具。<https://sites.google.com/site/cocotesttool/>

- PEX 微软针对 .NET 的白盒测试开源框架。结合了符号执行与搜索技术。可以作为 Visual Studio 的插件，自动生成高覆盖率的测试数据。

<http://research.microsoft.com/en-us/projects/pex/>

- GenProg 利用遗传编程技术对 C 语言程序进行故障自动修复的工具。

<http://dijkstra.cs.virginia.edu/genprog/#problem>

- EvoSuite 针对 java 程序的测试数据生成工具。

<http://www.evosuite.org/>

- JavaPathFinder 对 java 程序进行执行、检验的工具，可用于探测多线程程序所有可能的执行路径。

<http://javapathfinder.sourceforge.net/>

- Milu 基于搜索的高阶变异测试 (Higher Order Mutation Testing) 工具。

<http://www.cs.ucl.ac.uk/staff/y.jiaMilu>

作者简介

李征 博士，北京化工大学信息科学与技术学院教授、博士生导师、教育部新世纪优秀人才计划获得者。中国计算机学会（CCF）高级会员，软件工程专委会和容错计算专委会委员，担任 STVR、JSS、JSEP 等国际期刊客座编辑，IEEE SCAM 2012 程序委员会主席，ICSM、WCSE、GECCO 和 RT 等多个国际会议的程序委员会委员，是中国基于搜索的软件工程研讨会（CSBSE）发起人。主要研究领域：基于搜索的软件工程、程序源代码分析，目前已发表文章被引次数超过 500 次。



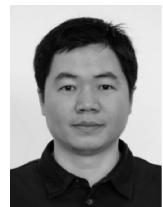
巩敦卫 博士，中国矿业大学教授，博士生导师，教育部新世纪优秀人才支持计划入选者，江苏省“六大人才高峰”高层次人才，江苏省“333 高层次人才培养工程”培养对象，全国煤炭青年科技奖获得者，IEEE 会员，中国计算机学会会员。研究方向为：基于搜索的软件工程、智能优化与控制，主持国家自然科学基金 5 项、国家“973”计划子课题 1 项，研究成果获省部级科技奖励 6 项。



聂长海 博士，南京大学计算机科学与技术系教授，博士生导师。中国计算机学会（CCF）高级会员，软件工程专委会委员。一直从事软件工程，特别是软件测试领域的教学和科研工作。发表软件测试学术论文 70 多篇，包括软件工程领域 A 类期刊 TOSEM 及计算机领域最具影响力的期刊 ACM Computing Surveys。主讲软件测试等 10 多门课程，独立编写并正式出版了《软件测试的概念与方法》，获得过发明专利和多项软件著作权，主持完成过国家自然科学基金项目、科技部 863 专题项目和江苏省自然科学基金项目，参与的项目多次获得过省部级科研奖。



江贺 博士，大连理工大学软件学院教授、博士生导师、教育部新世纪优秀人才计划获得者。Applied Intelligence Journal 客座编辑，Frontiers of Computer Science 青年编委，中国计算机学会软件工程专委会，中国计算机学会计算机应用专委会委员。主要研究领域：基于搜索的软件工程、软件仓库挖掘。先后在 IEEE Transaction 系列汇刊（TSE, TKDE, TSMCB, TCYB）、EC、中国科学等期刊及 ICSE、GECCO 等国际会议发表论文 60 余篇。



计算机辅助设计与图形学研究进展与趋势

CCF 计算机辅助设计与图形学专业委员会

鲍虎军¹ 陈为¹ 冯结青¹ 刘利刚² 王锐¹ 张松海³

¹浙江大学 CAD&CG 国家重点实验室，杭州

²中国科学技术大学数学科学学院，合肥

³清华大学计算机系，北京

摘要

计算机辅助设计与图形学是计算机科学领域的一个重要分支。基于计算机辅助设计与图形学的理论和方法的一系列高技术产品和系统，在一定程度上改变了人们的工作和生活方式。产业和市场需求的驱动、硬件技术的发展，促使计算机辅助设计与图形学的研究人员不断探索新的应用领域、新的理论和方法。本报告介绍了近两年来计算辅助设计图形学领域的研究热点和难点，重点介绍了图形绘制、数字媒体、可视化与可视分析以及三维打印四个研究方向的国内外研究动态、研究成果，并进行了分析和对比，最后对中国的计算机辅助设计与图形学发展进行了展望。

关键词：计算机图形学，计算机辅助设计，图形绘制，数字媒体，可视化与可视分析，三维打印

Abstract

Computer Aided-Design and Computer Graphics (CAD&CG) is one of the most active research streams in the development of computer science and technology. Various high-tech products and systems based on CAD&CG come into the daily life and impact people's work and life styles. On the other hand, industrial and market demands and graphics hardware developments facilitate the researchers to explore new application area, novel theory and methodology. This report briefly introduces the major progresses of the Chinese researchers on graphics rendering, digital media, visualization and visual analytics, 3D printing with focuses on novel algorithm design, new technology development and successful applications. The state of the art in the above fields is summarized and the future R&D directions are suggested.

Keywords: computer graphics, computer- aided design, graphics rendering, digital media, visualization and visual analytics, 3D printing

1 引言

计算机辅助设计与图形学经过 40 多年的发展，传统的研究方向（例如图形绘制）正

在面临一些瓶颈问题的挑战；同时，随着硬件和网络技术的飞速发展，又为计算机辅助设计与图形学的研究提供了新的发展机遇与挑战，例如数字媒体、可视化与可视分析、三维打印等。

图形绘制研究是计算机图形学的重要研究内容之一，也是图形学研究领域的重要课题。经过几十年的发展，已获得丰硕的研究成果，大大提高了计算机对虚拟世界的呈现能力。然而，随着人们对客观世界的认识进一步深入，计算机图形技术的普及和计算机图形技术应用的进一步发展，人们对绘制结果的高真实感与绘制过程的快速性要求变得越来越高，这给图形绘制研究提供了源源不断的发展动力与目标。近年来的图形绘制研究，主要呈现出以下几个特点：“新”——图形绘制研究不断取得突破，模拟出新的、原来难于模拟的绘制效果，让计算机模拟的虚拟世界更加真实；“精”——在图形绘制几十年研究积累下来的许多算法（例如路径跟踪、辐射度方法）近年来都获得了进一步的发展，原有方法的桎梏被不断突破，新的应用领域被开拓，从而诞生出比原有方法更精密的绘制算法；“尖”——新的硬件、新的计算架构、新的理论框架被引入到绘制研究中，从而引领了一批最尖端的绘制研究成果。

数字媒体是指以二进制数的形式记录、处理、传播和获取信息的载体，这些载体包括数字化的文字、图形、图像、声音、视频和动画等媒体。在当前爆炸性的数据增长中，数字媒体已成为最主要的数据来源，占到数据总量的 80% 左右。美英日等发达国家数字媒体产业规模在本国 GDP 均占有重要份额，数字媒体技术体现了一个国家在信息服务、前沿技术研究和集成创新方面的实力和产业水平，并成为目前学术界研究和产业界研发的热点方向。

数据可视化和可视分析作为一个新兴的研究领域，受到越来越广泛的关注。在麦肯锡 2011 年发布的一个报告^[1]中，可视化被列为基于数据创新的关键技术之一。数据可视化公司 Tableau^①于 2013 年 5 月在美国纽约股市挂牌，成功融资 2.54 亿美元，市值达到 20 亿美元。美国科学院 2014 年发布的一本关于大数据分析前沿的调研报告^[2]中，可视化被认为是一种混合式人机融合的数据分析技术，“不仅帮助人类理解分析的输出，同时提供用户修改数据分析模型的手段”。

三维打印是增材制造技术（Additive Manufacturing, AM）的俗称，它依据三维 CAD 设计数据，采用离散材料（液体、粉末、丝、片、板、块等）逐层累加制造物体的技术。相对传统的材料去除方式、材料成型方式，三维打印是一种自下而上材料累加的制造工艺，自 20 世纪 80 年代开始逐步发展，也被称为快速成型（Rapid Prototyping）、分层制造（Layered Manufacturing）等。它改变了传统的减式材料制造模式，带来了制造工艺和生产模式的变革，有力推动了三维数字化相关技术与研究的发展。在三维打印中，三维模型是前提和基础，三维打印是结果，它使三维模型“落地开花”。但是，大多情况下，现有方法直接得到的三维模型并不能直接输出给三维打印机。因为大部分设计模型都是由建筑师、工程师或设计人员所提供，他们都倾向于使用专业设计软件，如 Maya、

① <http://www.tableausoftware.com/>.

3ds Max 和 SketchUp 等。还有一些三维模型数据来自于三维扫描设备，如激光扫描仪、结构光扫描仪等。这些模型数据信息并未考虑到三维打印的具体需求与约束，如果直接输出到三维打印机，通常会导致各种各样的问题，如可能模型尺寸过大，超过打印机能打印的尺寸限制或没有考虑稳定性导致打印出物体无法正常放置等。因为以上原因，大多数设计模型，尤其是那些复杂物体的三维模型，都需要经过一些计算机图形学方法进行修正、调整和优化，使其能更好地满足三维打印的需求，避免打印出的物体无法正常发挥功能。因此，近年来，三维打印中的计算机图形学研究工作得到了更加广泛的关注，三维打印中的几何处理工作成为研究的热点。

2 国际研究现状

2.1 图形绘制

图形绘制技术是计算机图形学的传统研究内容，经过几十年的发展，涌现出丰硕的成果。随着新的应用需求的发展，图形绘制依然是计算机图形学研究的热点领域。在 SIGGRAPH 和 SIGGRAPH ASIA 的会议上，每年都有 12~20 篇绘制研究相关的论文。这些研究覆盖了绘制研究的各个方面，这其中如下几个方向近年来获得了持续的关注，并取得了重要的进展。

(1) 全局光照绘制

高真实感的效果生成，需要模拟光线在含有大量细节的三维场景中的真实传输。由于光线传播的全局性，当模拟光线在场景中的多次反射、折射所产生的全局光照效果时，往往需要很大的计算代价。因此，研究快速、高效的全局光照算法一直以来都是真实感绘制研究的重要方向。在这一方向上，国内外的研究学者做了诸多努力。

首先，研究人员针对传统绘制方法的缺陷，提出了一系列改进与更新的方法，获得了更好的结果、更快的绘制速度。在路径跟踪上，研究人员提出了新的分治方法^[3]、利用低秩性假设^[4]、基于光线梯度变化^[5]等多种加速方法来提高绘制的效率。在光子跟踪上，人们提出了自适应的渐进式光子跟踪^[6]、利用光子可见性的自适应光子跟踪^[7]等方法来加速光子跟踪方法。在此基础上，人们进一步将路径跟踪与光子跟踪相结合，提出了统一的路径跟踪方法^[8,9]。此外，研究人员进一步扩展了多光源 (many-lights) 中基于光割的方法，将其扩展到双向光割^[10]以及超大规模场景的全局光照绘制之上^[11]。

在改进传统研究方法之外，一个新的研究趋势是通过对计算全局光照所需要的光场信号进行频谱分析，来得到光线在不同场合、不同应用中传播的特性，从而能够制定相应的光线采样与重构方法，进而加速绘制方法。研究人员^[12]通过对简化假设（二维）情况下的全局光照理论分析来帮助构建高效的三维全局光照方法。通过对光场信号在入射

辐射亮度场^[13]、软影^[14]、景深以及运动模糊^[15]等应用场合下传播的频谱分析，研究人员定义了对应的滤波函数来对采样光线进行重构，从而获得了更好的绘制结果。此外，人们也创造性地提出了多种滤波函数^[16,17]，进一步提高了光线重构的效率，增强了绘制结果。

(2) 复杂材质的表示与呈现

真实世界复杂多样的物体表面呈现出颜色各异的材质属性，为了模拟光线在这些不同的物体表面上的作用效果，科研人员从材质的表示模型、绘制算法等角度入手，开发出不同的复杂材质的呈现方法。这些复杂材质的真实呈现，大大加强了计算机模拟的虚拟世界的真实感。随着获取手段的丰富与应用需求的增长，复杂材质的表示与呈现方法是近来绘制研究的热点。

在材质的表示模型上，研究人员提出了一系列新的模型，从不同的光路分析、新的数学形式以及通过引入人类感知等不同的角度构建新的材质模型的数学表示，这些成果包括：基于折射的光反射模型^[18]、基于各向异性高斯函数的表面材质模型^[19]、基于稀疏采样重构的表面材质模型^[20]、基于视觉感知的半透明散射相函数模型^[21]等。在材质的绘制方法上，研究人员则从新的数学逼近方法、加速结构以及新的绘制策略等方面提出了一系列新的方法。研究人员提出了基于解析双重积的高光表面快速绘制方法^[22]、基于双尺度的高光函数编辑与绘制方法^[23]、基于虚拟线光源的半透明材质绘制方法^[24]、基于数据字典的半透明材质逆向绘制方法^[25]等一系列方法，进一步提高了计算机模拟复杂材质绘制效果的真实度、加快了计算模拟与绘制的速度。

(3) 新的绘制架构、硬件加速方法与高效采样方法

几十年的图形绘制研究构建了一系列的绘制架构用于真实感绘制、实时绘制，例如广泛用于电影绘制的 REYES 框架、用于基于物理绘制的路径跟踪框架、用于实时绘制的前向绘制框架与推迟渲染框架等。更好更快的绘制框架，一直以来都是研究人员的目标。近年来，在绘制框架的研究上，一方面研究人员尝试研究新的绘制框架，例如基于排序的推迟渲染框架^[26]；另一方面，研究人员更多地将通用计算引入到传统的绘制框架里，将多种绘制框架相结合，尝试利用硬件来并行加速计算，提高绘制效率^[27,28,29]。

此外，针对不同绘制方法所共有的一些基础问题，例如用于离散积分的蓝噪音采样^{[30][31]}、针对可见性的采样方法分析^[32]等，科研人员也开展了一系列研究，获得了新的结果。

2.2 数字媒体

数字媒体技术主要研究与数字媒体信息的获取、处理、存储、传播、管理、安全、输出等相关的理论、方法、技术与系统，其所涉及的关键技术和内容主要包括数字信息的获取与输出技术、存储技术、处理技术、传输技术等。数字媒体处理技术涉及图像处理、计算机图形学、机器学习和认知分析等多个领域，在多个国际著名会议和期刊上都得到了广泛的关注，会议包括 ACM SIGGRAPH、ACM Multimedia、CVPR、ICCV 和 ACM

CHI 等, 期刊包括 ACM Trans. Graphics、IEEE Trans. Pattern Analysis and Machine Intelligence、IEEE Trans. Visualization and Computer Graphics、IEEE Trans. Multimedia 和 ACM Transactions on Computer-Human Interaction 等。近两年来, 数字媒体技术的研究热点主要集中在以下几个方向:

(1) 高维媒体(如 RGB-D 图像/视频、双目图像/视频、多视点图像/视频、光场、真三维图像/视频)的获取、分析与处理

传统的图像/视频是三维物理世界的二维映射, 存在深度信息的缺失。这种重要信息的缺失使得传统图像视频的处理存在很大困难。带有深度或多视点的图像/视频承载了更丰富完善的真实场景信息, 随着深度恢复、三维重建等技术的不断进步, 高质量采集技术和设备逐渐普及, 针对此类高维媒体分析与处理的研究成果快速增长。在 2013 年 Siggraph、Siggraph Asia、CVPR 和 ICCV 中均有三维重建、计算摄影学、多视点视频处理、光场编辑等相关工作的小节, 2014 年的 Siggraph 中更有超过十个工作与高维媒体的获取与处理相关。

(2) 利用海量数据的媒体信息智能处理

极度增长的海量数据为媒体信息处理提供了丰富的素材和知识, 为媒体信息的智能处理提供了难得机遇; 同时, 海量数据也导致信息过载, 使人们容易迷失在信息海洋中反而无法获取所需的素材。如何有效组织互联网的数字媒体素材, 及时准确地搜索到有用信息, 提供数字媒体的智能应用, 在近两年是一个研究热点。在 Siggraph 2013 和 2014 中, 均有超过十个工作与此相关, 包括图案配色^[33]、辅助作画^[34]、卡通肖像画创作^[35]等。

(3) 人体的图像/视频分析与重构

人对场景中的人及其行为非常敏感, 人物表情、手势、姿态是交流沟通的重要信号, 因此人的数字信息是数字媒体中一类非常特殊的信息, 包括脸、手、人体、个体和群体行为动作等。由于人对此类信息的认知敏感性, 人的图像/视频分析的质量要求非常高, 也具有特别的重视度, 近两年在 Siggraph、Siggraph Asia、CVPR、ICCV 也有专门 session 对此类工作进行论述, 主要集中在人脸表情跟踪与合成^[36,37]、基于视频的人手运动捕捉^[38]、人体^[39]运动识别^[39]、行为检测与分类^[40,41]等。

2.3 可视化与可视分析

2013 年, IEEE 可视化周 (VisWeek) 更名为 VIS, 分别代表其中三个最重量级的会议, 可视分析 (VAST)、信息可视化 (InfoVis) 和科学可视化 (SciVis)。除了这三个会议外, 2013 年的可视化周还包括了大数据可视分析 (LDAV) 和生物数据可视化 (BioVis) 两个会议, 涵盖体育数据、移动数据、医疗数据等方面的研究会, 以及以如何评估可视化为话题的专家讨论环节。

科学可视化是解决自然科学研究中前沿科学问题必不可少的数据分析工具^[42]。当前, 科学可视化研究的重点是三维空间物理化学演化规律、生物医疗等数据的可视化。

最近，不少工作将广义信息可视化的方法引入到科学可视化的研究中，如基于信息熵的多变量体数据的可视化^[43]。随着数据复杂度的提高和规模的扩大，高性能可视化方法也是研究的热点，工作集中在大规模空间流场和张量场的可视化和分析，解决问题的着眼点更是推广到 I/O、数据管理、工作流等方面^[44]。

信息可视化为理解和诠释各种海量复杂信息提供了手段^[45]。美国在“9.11”以后，出于反恐涉及的大规模复杂数据分析的实际需要，在全国范围内组建了以全美多所大学为地区分中心的国家可视分析研究体系。欧洲也紧随其后，由横跨欧盟十余个国家的数十个研究机构协作联合开展欧盟框架下的可视分析研究。与人机交互、用户测试紧密结合是信息可视化的重要特点。多维数据可视化和可视化用户测试一直是信息可视化的重要研究方向。针对复杂图、日志、社交网络等数据类型的可视化是新的研究热点^[46]。

可视分析作为新兴的可视化方向成为三个会议中最活跃的一个。2013 年的研究工作充分展现了可视化在其他领域中的作用。该年的研究工作渗入线性回归分析^[47]、时序数据的模型选择^[48]和高维数据投影的可解释性探索^[49]等传统数据挖掘领域；在文本、图像、视频以及社交媒体中的比较、聚类和交互式分析^[50~52]也有非常多的研究成果。可视分析的另一个重要方向是可视推理。

从这几年可视化周的发展以及收录论文的话题来看，可视化经过最初的探索和积累，逐渐形成了一个成熟的学科。在算法与技术之外，可视化处理的数据趋于复杂化、趋于大规模；可视化学者不仅关注技术的使用，对于可视化效果的评估和可视化的办法论也越加重视，逐渐形成了可视化评估和可视化方法论的理论体系。如 Matthew Brehmer 和 Tamara Munzner^[53]将可视化中的任务抽象为多层次的 3 类 7 种的任务组合，用于指导可视化的设计和可视化效果的评估。

2.4 三维打印

三维打印是近几年引起广泛关注的领域，其中对于三维模型的设计与处理的相关研究在计算机图形学和 CAD 领域涌现出大量的研究论文。

(1) 几何优化问题

三维打印中需要对三维几何模型进行优化，使得几何模型能够满足三维打印的需求。

1) 物体分割

一台三维打印机可打印对象的最大尺寸因为三维打印机本身空间有限而受限。因此，打印一些大体积的物体，对现有的三维打印技术而言，仍困难重重。对一个超过可打印尺寸的大物体对象，如果要将其三维打印，一个可行的解决方案就是将其分割为一块块可打印的小对象，然后再将其组装成一个整体大物体。针对这一问题，Luo Linjie 等给出了一个名为 Chopper 的分割处理方案^[54]。该方案采用平面分割，自上而下，每次分割均将处理对象一分为二，逐步细化，最终整个模型可形成一个 BSP 树的层次分割结果。对此问题，Chen Desai 等则给出了一个近似表示的方案^[55]：将一个三维模型转化为分片多边形面片近似表示，再通过三维打印每一个多边形面片，最后将这些面片拼装成一个与

原三维模型相近似的实物对象。

2) 重心优化

在三维虚拟环境下，三维模型可以任意摆放位置与姿势，包括可摆出违反重力原则的造型，因为在虚拟世界中，三维模型无需遵循真实世界中的物理规律。但是，如果把三维模型打印输出为实物，这时物理规律就要发挥作用了，如果它在各种受力情况下不能保持稳定状态，那它就不能很好地摆放到所需的状态。针对这一问题，Romain 等给出了重心优化方法^[56]，即通过几何方法来优化模型的重心位置使其在给定姿势下达到平衡状态。

(2) 结构分析问题

三维打印技术促进了产品个性化定制的普及与推广，使得每个人都可以设计三维几何模型，成为自己产品的设计师。他们由于缺乏一些设计经验与力学知识，会导致其设计结果直接三维打印后存在一些结构问题，如强度问题、稳定性问题等。强度不足可能会使三维模型在打印、运输或日常使用过程中受到破坏，而稳定性问题则会导致三维模型无法正常地放置或悬挂，影响其日常使用功能。这种问题我们称其为结构分析问题，它的主要任务是识别三维模型中存在的强度或稳定性缺陷，并给出适当合理的弥补方案。

针对强度问题，Stava 等给出了一个自动检测并修正结构强度问题的系统方案^[57]，来创建一个新的三维模型，使其与原有模型保持尽可能相近的外形，同时提高其结构强度与整体性。而 Zhou Qingnan 给出了一个更好的方案^[58]，该方案在预测或检测模型结构强度问题时，与上述明确指定或设定模型的荷载情况方法不同的是，它会去寻找一种最不利荷载情况（Worst-Case），并据此识别出模型上最易破坏之处或最大变形区域。

(3) 材料表面效果定制

随着可供三维打印材料类型的增多，人们希望能打印出更复杂外观、更多表面光学特征及力学特性的物体。这一需求催生了三维打印中一类重要但尚未很好解决的问题：如何确定出一个物体对象的材料组成，使其能满足一个给定的表面外观效果或变形功能要求。这一问题可称为材料表面效果定制问题。近年来，很多学者对此问题做了深入研究，其工作大致可分为以下三类：

1) 次表面散射效果定制（Subsurface Scattering）

为了使三维打印结果具有指定的次表面散射效果，Hasan 等采用 BSSRDF 函数来确定材料的次表面散射特性，并给出了一个完整的流程^[59]，包括测量、预估、匹配计算、优化和输出等步骤。Papas 等研究了通过不同的颜料与基本原料相混合来实现给定材料次表面散射效果^[60]。

2) 空间变化反射效果定制（Spatially Varying Reflectance）

在多数情况下，真实世界物体同一种表面反射效果还会随视角空间方向变化而变化。在计算机图形学中，常用双向反射分布函数（Bidirectional Reflectance Distribution Function，BRDF）来表示这种空间变化反射效果。自然，在三维打印中也会考虑如何打印出指定空间变化反射效果。为了定制出期望的表面外观反射效果，Weyrich 等结合 BRDF 函数，给出了一个基于微平面（Microfacet）理论的系统方案^[61]。后续相关研究也

大多通过 BRDF 函数来实现上述效果定制^[62~64]。

3) 变形效果定制

不同材料组合的联合打印可以消除传统单一材料打印的不足与局限，使我们能够制造更加复杂的物体，甚至能使这些多元材料转化为复杂的、新的功能材料，如同时兼具轻质和高强度性能的材料，或同时具备良好柔韧性和透明效果的材料等。Bickel 等就研究了上述这一很有实用价值的材料混合问题^[65]：如何在微尺寸的尺度（也即三维打印的尺度）上，根据基础材料的力学性能，打印出指定力学性能的基础材料组合体。

综合以上的定制处理方法，Chen Desai 等发现上述处理过程存在一些类似的流程与相同的处理单元，如它们都依赖于在给定几何与材料要求下精确模拟所给对象物理特征的能力。因此，文献^[66]中提出一个更具普适性的定制框架来处理上述问题，该框架具有模块化、可扩展性、打印设备无关性与模型几何无关性的特点，并给出了一些定制效果。

(4) 机构设计

三维打印不仅可以输出复杂模型，同时还能实现以往能设计但很难制作的机构。因此，最近两年各种机构设计的研究越来越多。这方面的研究主要可分为两大类。一类是静态机构设计，如积块式机构设计，这类机构的构件按一定方式组装起来，形成一个稳定的形状，如鲁班锁^[67]、联锁积木^[68]、Schwartzburg 和 Pauly 的交错式片块机构^[69]。另一类是动态机构设计，这类机构可以活动或运动起来。具体包括：a) 动态玩具机构，如机械角色设计^[70]和机械人设计^[71]等；b) 关节机构，如 Jacques 等^[72]和 Moritz 等^[73]这两篇文章均是针对有关节的角色模型对象，虽然实现细节略有差异，但都实现了角色关节机构的免组合安装（Non-Assembly），很好地体现了三维打印一体成型的特点和优势。

(5) 自支撑结构设计

近年来，用三维打印来设计验证建筑或结构设计的研究也越来越多，自支撑结构设计便是其中的代表。如何在数字世界与真实世界中设计出指定外形的砖拱结构，仍是一个既具有重要价值又具有一定难度的任务。现有方法大多是基于推力网络分析方法（Thrust Network Analysis, TNA）^[74]来处理的。这种方法需要有深厚的结构设计知识基础，与大量的人工设计计算工作。在这样的背景下，近年来有一些研究者开始探讨如何利用计算机来处理上述问题，如 Daneile 等^[75]，Fernando 等^[76]，Vouga 等^[77]。自支撑结构中，还有一类 RF 结构（Reciprocal Frame Structures）也很有价值，这种结构由一些 RF 单元互相搭接而成。如何利用计算机的方法来完成复杂 RF 结构设计，就变得很有必要，Song Peng 等^[78]就对此问题做了一些探讨。

(6) 内部结构设计

1) 编解码嵌入技术

编解码嵌入技术可以在产品生产过程中把多种格式的各种信息插入到一个物体内部，为这些产品包含身份信息标记。Wilson 和 Willis 提出了一项在三维打印中把信息嵌入到物体中的技术^[79]。这项技术称为 InfraStructs，它是将太赫兹（THz）扫描嵌入到三维打印中，在三维加工的过程中可将编码信息隐藏，随后解码成为有效标签信息。

2) 多层模型结构

水晶内雕是一种颇为引人注目的工艺品，它是在水晶、玻璃等透明材料内雕刻平面或三维立体图案，如可雕刻 2D/3D 人像、人名手脚印、奖杯等个性化标志信息，可呈现出立体逼真、光彩夺目的效果。能否用三维打印来生成上述水晶内雕的效果呢？Michael 等研究了通过多层结构模型来快速打印生成类似效果的物体^[80]。

3 国内研究进展

3.1 图形绘制

图形绘制是图形学研究的传统方向之一，在国内主要开展绘制研究的是浙江大学、清华大学、中科院软件所以及微软北京亚洲研究院等高校和科研单位，在 2012~2013 年期间共在 ACM TOG（包含 SIGGRAPH, SIGGRAPH Asia）上发表论文 5 篇，IEEE TVCG、CGF 等国际期刊上发表论文 8 篇。国内学者在图形学绘制领域的工作主要围绕全局光照绘制、复杂材质呈现、高效采样方法等方面开展研究。

(1) 全局光照绘制

在全局光照绘制方面，国内的研究学者紧跟国际研究的潮流，开展了多项研究，在全局光照的快速计算、超大规模场景的全局光照计算以及阴影效果的快速绘制上取得了一系列成果。

针对复杂光线传输，清华大学与微软北京亚洲研究院的任沛然等研究人员^[81]在 SIGGRAPH 2013 上提出了采用回归函数来逼近静态场景中辐射亮度传输，进而实现对三维场景实时全局重光照绘制的技术（Global Illumination with Radiance Regression Functions）。该技术将机器学习和图形学相结合，通过回归分析与预计算，将辐射亮度在三维场景中的传输作为训练数据，学习与训练存储在三维场景上的神经网络基函数，从而获取三维场景对辐射亮度传输的降维与简化表示。由于神经网络学习方法的高适应性与非线性性，相比前人的方法，该技术可以高效地表示与逼近复杂的辐射亮度传输函数，例如镜面反射产生的焦散效果、高频的高光反射等，并能在绘制时实现对光能的实时重构。但由于需要预处理过程对光能传输进行计算并进行回归计算，该技术仅支持静态场景且需要较长的预计算时间。

针对大规模复杂场景，浙江大学的王锐等研究人员^[11]在 SIGGRAPH ASIA 2013 上提出一种基于外存的 GPU 加速多点光源绘制技术（GPU-based Out-of-Core Many-Lights Rendering）。大规模场景的全局光照绘制一直以来是图形学绘制研究的难题。特别是当场景规模超过内核内存（in-core memory）的容量，需要采用基于外存的存储（out-of-core）方式时，绘制计算与数据调度的效率成为制约绘制计算的重要瓶颈之一。基于近年来发

展起来的多点光源绘制框架 (many-lights rendering framework)，该技术提出了基于 GPU 的大规模场景多点光源绘制方法。该方法通过将多光源的积分计算分解为光传递子矩阵，充分利用了当前 GPU 的海量并行能力来提高计算的效率。为解决超过显存的光源数据几何数据的调度问题，该方法将采样光传递矩阵的过程看做一个货郎担的全图遍历问题。通过对该货郎担问题的优化求解，实现了在有限计算时间内的高效数据调度。包含上亿面片与数千万至上亿的点光源场景，该方法可以取得比 CPU 计算一个量级的绘制效率的提高。

物体对光源遮挡所产生的阴影效果，作为全局光照效果的一种，可以给人们提供重要的物体深度与相对位置关系的线索，因此获得了研究人员的重点关注。如何快速而且便捷地生成阴影长久以来都是研究的热点。浙江大学的王锐等研究人员^[82]在《中国科学》上提出了一种采用阴影几何图的阴影绘制技术。该技术针对传统阴影图算法由于阴影图采样精度的限制，生成阴影边缘具有明显走样的缺点，通过在传统的阴影图上存储额外几何信息，实现物体间精确的光源遮挡计算，从而生成无走样的阴影效果。浙江大学的沈笠等研究人员^[83]在 EGSR2013 上提出了基于指数阴影图的软影绘制技术。该技术将针对点光源阴影的指数阴影图技术扩展到了针对面光源的软影计算上，通过充分利用指数阴影图高质量的滤波重构、低空间存储与低计算消耗等优点，实现了高质量的软影绘制。

(2) 复杂材质呈现

国内的研究人员在此方面也开展了多项研究，在材质函数的表示、复杂材质的快速绘制与呈现、半透明材质物体的绘制与编辑以及毛发材质的真实绘制上取得了一系列的成果。

针对复杂表面材质在复杂光照（例如环境光）下的实时呈现，浙江大学的王锐等研究人员^[22]在 IEEE TVCG 上提出了一种双解析积分的方法用于快速计算绘制方程的积分。在只考虑环境光照的前提下，传统上绘制方程被表示为环境光、物体表面材质函数与可见性的三重积分。然而，计算该三重积分需要较高的计算代价，在模拟全频段光照效果时，无法达到实时的绘制速度。针对这种情况，该项技术将三重积分转化为在可见区域上的双重积分，通过解析的二重积分计算来加速光照计算。该项技术利用了前人提出的球面高斯函数的材质表示方法，并采用自适应距离场函数来划分可见区域，此外，通过充分利用 GPU 的并行性，该项技术可以在环境光下实现复杂材质的全频段光照呈现，并达到实时的绘制速度。随后，针对该方法只能处理静态场景的限制，王锐等研究人员^[84]在该计算框架下进一步提出了在线计算可见性的方法，将上述技术推广到动态场景，实现了中等规模动态场景的全频段复杂材质呈现。

清华大学的徐昆等研究人员^[19]在 SIGGRAPH ASIA 2013 上提出了一种各向异性球面高斯函数 (Anisotropic Spherical Gaussians)。此项工作在使用各向同性的高斯函数来表示物体表面材质函数的基础上，进一步提出了使用各向异性的球面高斯函数。该函数的参数可以适应各种不同形状的各向异性的分布，例如铝的表面材质特性，大多都是各向异性的，即其沿垂直于轴向的截面都是椭圆。基于此表达，很多真实材质分布函数只需

要一个各向异性球面高斯函数就可以很好地表达。同时，各向异性球面高斯函数保持了传统球面高斯的乘法封闭性，且拥有近似的积分解析解。利用新函数的特性可以将拥有各向异性材质的场景的渲染效果和帧率大幅度提高。

清华大学的闫令琪等研究人员^[85]将球面高斯函数推广到半透明物体材质的呈现中，在PG 2012上提出了基于球面高斯光的半透明材质绘制技术（Accurate Translucent Material Rendering under Spherical Gaussian Lights）。该方法充分利用了球面高斯函数乘法封闭和积分具有解析解的数学特性，通过用一个或多个球面高斯函数近似表达半透明材质中的光照分布从而利用球面高斯函数的特性实现绘制的加速计算。针对半透明材质，浙江大学与微软北京亚洲研究院的Dongping Li等研究人员^[86]在IEEE TVCG上提出了一种采用有限元方法对散射方程进行快速求解的技术（TransCut: Interactive Rendering of Translucent Cutouts），利用该技术可以实现对光能在半透明材质中的散射进行快速计算，从而实现对形状变化的半透明材质进行快速呈现。

针对其他更为复杂的表面材质（例如毛发），浙江大学的秦昊等研究人员^[87]在IEEE TVCG上提出了基于光锥跟踪的绘制技术（Cone Tracing for Furry Object Rendering）。传统光线跟踪只能跟踪一根光线，而这项技术针对毛发材质对光的反射特性，采用了光锥来计算较远方向上毛发的着色、透明与遮挡效果，从而大大提高了绘制速度，并且能够方便地支持景深、运动模糊等特效。

（3）高效采样方法

高效的采样方法直接影响了绘制积分计算的效率与质量。在多种采样方法中，人们常采用蓝噪声采样。蓝噪声（Blue-noise）采样是指采样生成随机且均匀分布的采样点集合，由于其生成的随机样本的高均匀性，该方法已被广泛用于真实感绘制的光场函数、BRDF等高维函数的采样中。国内的研究人员在提高蓝噪声采样适用范围与采样效率上也开展了一系列工作。

微软亚洲研究院与浙江大学的孙鑫等研究人员^[31]在SIGGRAPH 2013上提出了一种生成蓝噪声线段采样的技术。传统的蓝噪声技术主要考虑空间点的采样，而在景深、运动模糊、散射介质的路径跟踪绘制计算中，往往需要随机生成线段采样。该技术通过对线采样的频谱分析，得到了线采样与点采样的对应关系，从而可以通过调制点采样与线采样在采样噪声与采样走样之间得到平衡。该方法被应用于景深、运动模糊以及散射介质的绘制中，获得了比传统方法更好的绘制效果，生成结果的噪声更小。

清华大学的陈家挺等研究人员^[30]在SIGGRAPH ASIA 2013上提出了一种双边蓝噪声采样技术（Bilateral Blue Noise Sampling）。传统的蓝噪声技术主要考虑采样点的空间属性，而不能鲁棒地处理更加普遍的采样点具有非空间特征的情况，例如光子映射中光子具有的光通量和入射方向等属性。受到双边滤波的启示，该技术的核心思想是引入采样点之间非空间特征的相似度，来调制传统的采样点间距离度量，这样双边蓝噪声分布既能反映采样点空间位置的均匀性，又能体现采样点非空间属性的特征。该文对双边蓝噪声采样进行分析和合成，并将其运用于物体分布、光子密度估计和点云欠采样等图形应用中。

厦门大学、香港大学与中国科技大学的陈中贵等研究人员^[88]提出了一种基于变分方法的蓝噪声采样技术。通过将传统方法计算采样点间等距的 Voronoi 细分替换为变分优化，该技术可以更好地处理动态变化域上的采样，并能更好地保持采样点的时间一致性。

3.2 数字媒体

国内多家高校和科研院所在数字媒体获取、处理与传输方面开展了卓有成效的研究工作，包括浙江大学、清华大学、北京大学、中科院深圳先进技术研究院、中科院自动化所、中国科技大学、国防科技大学、山东大学、北京航空航天大学、中科院软件所等。

在获取方面，清华大学建立全光协同采集平台，在国际上首次提出视觉场的概念，解决了未知光照下自由运动对象的高精度立体建模难题，在光场采集和处理^[37]、立体视频重建^[89]等方面取得了国际领先的成果。北京大学、中科院深圳先进技术研究院开发了室外真实场景采集系统与平台，并在腾讯公司街景系统的数据采集中得到应用。

在传输方面，北京大学在视频编码与传输方面做出了突出成果，成立了 IEEE AVS 工作组，成功制定了 IEEE P1857 国际标准，并于 2013 年 6 月 4 日颁布实施，开发了 AVS + 3D 视频编解码系统，支持双屏、全高清等多种立体视频格式，支持高清视频的实时摄录播一体化处理过程。上海交通大学完成了超高清 4K 数字电视内容制作、超高清信源编码、信道调制解调、超高清信源解码显示全链路统硬件平台，是我国首个 4K 超高清电视的完整链路平台，其系统技术已代表中国在 2013 年参与美国 ATSC3.0 国际数字电视标准的竞争，与高通、三星、SONY、LG 等全球知名企同台竞争，力争进入国际化标准体系。

在处理方面，数字媒体处理的需求非常广泛，数字媒体的智能处理体现在使用尽量少的交互量实现满足用户需求的数字媒体处理效果，这需要媒体内容的准确理解和用户意图的有效感知。在这方面，国内多个科研团队取得了创新成果，在国际顶级期刊和会议上发表了多篇论文。以 ACM SIGGRAPH 2013 为例，浙江大学参与了 4 篇论文的工作，清华大学参与了 10 篇论文的工作，国内其他高校与科研院所也有广泛参与，包括中科院深圳先进技术研究院、杭州电子科技大学、香港城市大学、华南理工大学、南京大学、大连理工大学等。ACM SIGGRAPH ASIA 2013 中，清华大学参与了 7 篇论文的工作，浙江大学参与了 3 篇论文的工作，其余参与的高校和科研院所所有中科院深圳先进技术研究院、香港大学、国防科技大学、中国科学技术大学、大连理工大学、香港中文大学等。这些工作涉及人脸、手、头发建模，三维模型库组织和场景构建，图像结构化表达和编辑，三维场景绘制等方面。取得的突出进展包括：

1) 中科院心理所和清华大学针对数字媒体处理的认知机理开展合作研究，提出了一种认知架构通用模型——PMJ 模型，将人类的认知机理归纳为一个可计算的认知阶段 - 通路框架，为视觉认知机理的计算建模提供了理论指导。在 PMJ 模型基础上，提出了一种面向物体识别的分布式计算认知模型，在归纳以往心理和神经生理学研究的基础上提

出了一个重要假设，即在认知加工中快速加工通路的感知阶段，人类视觉系统将对外部物理世界三维物体的视觉刺激感知为线条图的简明形式，再传递到后续的记忆和判断阶段。该假设在与中科院心理所合作进行的事件相关电位（ERP）技术研究中得到了初步的验证^[90]。基于此项研究发现，提出了基于线条图的局部编码表示、自适应学习和快速判断的计算机算法，并成功应用在图像、视频和三维数字模型等数据的高效智能处理中^[91]。

2) 在数字媒体的结构分析中，清华大学提出了一种图像结构化表示的“片网”模型^[92]。“片网”是一种表示图像层次结构的图模型，该模型利用图像中的区域具有相似表观特征的特性，用一个代表性片来代表一个区域，并作为该图模型的一个节点，并且在相邻节点之间加入边来表达其相对位置关系，从而对一幅图像进行抽象的表示。在编辑和合成等应用中，可以通过“片网”这种结构，实现图像库中局部结构的快速搜索，实现高效的图像处理。

3) 在三维场景建模方面，清华大学提出了一种从输入草图自动生成三维场景的交互技术 Sketch2Scene^[93]，利用互联网上三维场景模型构建场景模型库；通过分析场景库中的语义单元，利用多草图间的语义关联关系，提出草图联合分析方法，实现二维草图到三维场景的自动转化，大幅提高可视媒体交互构建的效率。提出基于单张图片的三维交互建模方法——“3-Sweep”^[94]，对给定图片中的物体进行建模，用户只需特定的三笔画就可完成建模过程，并在模型表面自动完成纹理贴图。用户还可方便地对模型的部件进行拉伸、旋转等高层语义的操作。

4) 在多视点图像/视频的分析与处理方面，清华大学通过设计一系列基于视觉感知的实验，测量人对不同双目立体视觉刺激的反应，定量分析双目视差、运动速度和图像频率对于舒适程度的影响，提出双目舒适度计算函数。基于双目舒适度计算函数，结合视觉系统帧内和帧间内容的融合机制，针对真实双目立体视频提出一种计算舒适度的度量^[95]。该工作为双目视频的质量评定、编辑和制作提供了很好的依据。在此基础上提出了双目立体图像的视点编辑方法^[96]。

5) 清华大学开展了数字媒体交互与合成方面的研究，提出了基于混合域计算的边缘敏感的图像调节方法^[97]，保证达到目标整体效果的同时能尽量保持住图像的边缘细节信息，不在边界处产生光晕等瑕疵；提出了基于四元树定性分析的模型组织方法^[98]，最大程度地保持由四元树所嵌入的模型相似程度的拓扑关系；提出了面向编辑前后图像对的编辑历史恢复方法^[99]，支持几何变换、颜色调节等多种编辑操作。

3.3 可视化与可视分析

可视化和可视分析经历过一段低潮后，随着大数据时代的到来，重新在国内引起广泛重视。北京大学、浙江大学、清华大学、中科院软件所、中科院计算所、中科院超算中心、北京应用物理与计算数学研究所、天津大学、山东大学等单位都开展了卓有成效的可视化研究。

(1) 发展概况

北京大学提出了一系列高维数据可视化方法，在面向高维、交通时空的可视分析方法与系统方面做了较有影响的工作。浙江大学的研究人员在高维时变数据的语义理解可视化等方面做出了创新工作。中科院超级计算中心的工作集中在天文、生物、气候、地质学等领域的大规模科学计算数据的可视化与分析，在 GPU 机群上实现了宇宙结构形成的 TB 级数据的可视化。

全国各地还加强了在可视化方向的人才培养。浙江大学出版的“十二五”规划教材《数据可视化的基本原理与方法》已经成为全国各高校开设的信息可视化课程的标准教材，该教材还配备课件、视频案例等丰富的教学资源；专著《大数据丛书：数据可视化》出版至今短短几个月已经售出 3 300 余本。可视化前沿研究生暑期学校受到全国各地学生和老师的欢迎，浙江大学的可视化暑期研讨班参加人数达到 400 人。天津大学举办了第五届可视信息交流与交互会议（VINCI '13），《清华大学学报（英文版）》发表“可视化和可视化分析”专题（第 18 卷，2013 年第二期）。

(2) 理论成果

北京大学发布微博可视分析系列工具[⊖]，利用新颖的可视分析技术帮助用户深入挖掘微博事件、关键词、用户等的关系。针对数据探索，他们设计了基于所见即所得的理念设计局部可视化方法^[100]、基于社区结构的网络数据探索技术^[101]和基于子空间的高维数据探索技术^[102]。他们的交互拥堵可视分析工具^[103]能够处理大规模的交通轨迹数据，对时间、空间进行过滤，分析某一路段的拥堵方向和原因，对指导城市居民出行、城市规划、交通管理等都有重大意义。

浙江大学基于超图对小说等长文档进行可视化^[104]，剖析长文档之间的话题转换关系。他们的移动社交网络可视分析工作^[105]基于上下文迭代式挖掘社交网络中的社区关系，该方法比一般的社区发现工作有更高的精确性。在大规模计算集群的可视监控方面，他们的在线性能可视化系统^[106]能够对系统性能的日志流进行实时监控。

天津大学的 Cube2Video^[107]能够将视频流实时转换为全景图，TabuVis^[108]是针对多维数据可视分析的工具，Vis4Heritage^[109]用可视分析方法探索莫高窟壁画的退化问题，对文化遗产保护的时空可视分析技术有了重大突破。

香港科技大学为网络数据中的集合^[110]和大规模运动数据的互换模型^[111]设计了新的可视化方法。运用传媒的话题竞争理论，他们分析了新闻媒体言论、政治人物言论和草根言论（如微博）在话题传播中的作用和话题的竞争^[112]。

(3) 产业发展

以微软亚洲研究院、IBM 研究院等为首的跨国企业研究院持续加强信息可视化和可视分析方面的研究力度。国内的海云数据利用计算机图形图像处理技术，为不同行业客户提供基于数据交互可视化服务的整体解决方案。他们于 2013 年底又成立了图易实验室致力于数据可视化学术价值和商业价值的研发。此外，数据新闻在国内发展如火如荼，

[⊖] <http://vis.pku.edu.cn/weibova/>.

知名发布源包括如网易数读[⊕]、CIVN 中文信息可视化社区[⊖]等。

在淘宝发布数据魔方、数据指数、dataav 可视化库等数据产品之后，百度奋起直追发布了基于网页的 canvas 绘图库 zrender[⊕]，并基于 zrender 发布纯 javascript 图表库 echarts[⊖]，使得在网页端展示大规模数据图表成为可能。

3.4 三维打印

国内的许多研究工作者在三维打印方面也做出了许多研究工作。

(1) 几何优化问题

三维打印中存在着许多对三维几何模型的约束和优化的问题。

1) 物体分割。针对分割问题，中国矿业大学的 Hao Jingbin 等给出了一个基于曲率的模型分割方法^[113]。该方法首先对模型表面进行曲率分析，提取出模型的特征边，并据其构建特征环。以此为基础，在其中选择合适的特征环来将原模型分解为小而简单的子模型组合。这种分割方法的前提是模型表面具有明确的特征信息，因此该方法适用范围有限。

2) 打印成本优化。与传统制造所生产的产品相比，三维打印产品的成本仍相对较高。因此，如何能在不牺牲打印物体质量的前提下，通过优化模型来减少打印材料消耗，对于降低打印成本至关重要。为节省打印材料，受建筑工程中的桁架结构的启发，中国科技大学的刘利刚教授团队提出了一种基于“蒙皮 - 刚架”(Skin-Frame) 的轻质结构来解决材料优化问题^[114]，发表于 SIGGRAPH Asia 2013。这种结构能有效地降低打印材料成本，并使打印物体满足所要求的物理强度、受力稳定性、自平衡性及可打印性。最近，山东大学的陈宝权教授团队提出一种峰窝状结构来解决同样问题^[115]，发表于 SIGGRAPH 2014。

(2) 结构分析问题

结构分析中有一类逆向弹性形状设计 (Inverse Elastic Shape Design) 问题，即当设计弹性物体形状时，在设计一开始时就考虑物体在受力后的弹性变形，并将这一变形结果再反作用于物体的初始形状，得到期望的设计形状。浙江大学的周昆教授团队采用 Asymptotic Numerical Method 对这一问题做了一些研究^[116]，成果发表于 SIGGRAPH 2014。

(3) 材料表面效果定制

1) 次表面散射效果定制。针对次表面散射效果定制问题，微软亚洲研究院的 Dong Yue 等给出了一套基于 BSSRDF 函数方案^[117]。该方案在给定的材料次表面散射特性要求下，可以有效地计算出所打印物体的每层材料分布及其厚度。其中，所给定的材料次表面散射要求也是由 BSSRDF 函数来描述。Dong Yue 等将此问题称为材料映射问题

[⊕] <http://data.163.com/>.

[⊖] www.civn.cn.

[⊕] <http://ecomfe.github.io/zrender/>.

[⊖] <http://echarts.baidu.com/>.

(Material Mapping)，即给定一组基本材料及分布约束条件，计算出物体材料组合使其BSSRDF 符合所给曲线要求。

2) 空间变化反射效果定制。微软亚洲研究院的 Dong Yue 等基于 BRDF 函数研究了如何定制一幅具有空间变化反射效果的 HDR 图像^[118]，当观察者变换观察角度时，这幅图像可以逐渐由暗变亮。清华大学的 Lan Yanxiang 等基于 SVBRDF 函数 (Spatially-Varying BRDF) 研究如何定制出具有各向异性的空间变化反射效果表面^[119]。

(4) 机构设计

杭州师范大学的许威威教授等研究了玩具机构的设计^[120]，以玩具角色的指定运动为主，同时加上玩具角色的几何信息，与玩具下方的方盒尺寸为输入信息，从预定义的部件库中选取合适的部件将其组合，然后再优化这些部件的参数，使整个机构的运动输出与所给运动输入保持一致。

传统机械设计时，对机械机构来说，一般都需要两步完成模型构建。首先，创建机构的每一个组成构件；其次，再将所有的组成构件装配起来。三维打印情况下，这一过程可以省去第二步，得到免组装机构。华南理工大学的 Su Xubin 等就基于选择性激光融化工艺研究了免组装机构设计有关问题^[121]。

(5) 自支撑结构设计

微软亚洲研究院的 Liu Yan 等^[122]在 TNA 的基础上引入 Regular 三角化 (Triangulation) 方法，给出了两者之间的联系，最后利用 Regular 三角化方法来参数化生成所需要的自支撑曲面，这种方法可提供一种便捷的参数化方式来交互创建或编辑自支撑曲面。

4 国内外研究进展比较

4.1 图形绘制

在图形绘制研究领域，经过多年发展与追赶，国内的研究紧跟国际研究的热点，在某些方面已经达到国际先进水平。例如，清华大学与微软北京亚洲研究院的任沛然等研究人员^[80]在国际上首次将机器学习的方法引入到全局光照绘制中，为后续的研究开辟了一个新的方向；浙江大学的王锐等研究人员^[11]首次实现了超大规模（亿万面片）场景复杂光照下的全局光照绘制；国内科研人员在蓝噪声采样方面的多项研究也为该领域的最前沿研究成果。

但是与国外的研究相比，国内研究也存在诸多不足。首先，相比国外开展绘制研究的各大高校、科研单位与企业，国内的高校和科研单位研究力量相对薄弱，研究内容更多的集中于绘制算法与现有方法改进，对于更为基础的绘制框架、绘制硬件的研究相对较少，在研究的基础性上仍有一定差距。其次，国内的研究缺乏国内企业的参与，这使

得国内的科研人员在研究内容的选择上更多地基于国外的应用需求，造成科研成果无法直接与国内图形绘制领域的应用挂钩，研究成果转化相比国外还严重不足。

4.2 数字媒体

在数字媒体的获取方面，研究一直致力于更高效地获取更丰富的视觉信息，当前研究重点主要集中在光场采集，麻省理工学院、斯坦福大学、哥伦比亚大学、Adobe公司、微软亚洲研究院等科研单位占据技术优势，采用微透镜阵列、摄像机阵列等不同方案，国内清华大学、北京大学、中科院自动化所等科研团队也搭建了光场采集系统，开展了深入研究，在此方面国内与国外先进技术差距不大。在 Siggraph 2013 中，麻省理工学院发表了关于飞秒级摄影技术^[123]，可以跟踪光线传播过程，值得国内科研团队高度关注。

在数字媒体传输方面，我国在数字视频编解码核心技术方面较为薄弱，相关企业长期受制于国外企业和组织持有的标准化专利与技术。由北京大学主导制定的数字音视频编解码技术国家标准（AVS）及其面向三维视频的 AVS+ 标准，通过采纳公开技术和我国的自主创新技术，把握了技术主动权。目前国际标准化组织也正在策划多视点以及全视角视频编码标准的制定，将使编码技术与光学全息技术实现结合。

在数字媒体的智能处理方面，高效生成满足用户需求的媒体内容是智能处理的最终目的。麻省理工学院、斯坦福大学等国外著名科研团队已经开展对数字媒体的认知机理研究。从人的认知入手，建立可实现的认知计算模型，将有可能带来媒体信息处理技术突破性的进展。同时，利用海量媒体数据作为支持，寻找其内在规律，研究海量数据驱动的智能处理方法，也是实现数字媒体的智能处理的有效途径之一。

4.3 可视化与可视分析

在可视分析领域，和研究最发达的欧美国家相比，我国无论是从事相关研究的科研机构的数目还是研究涵盖的方向，都还有一定的差距。在可以预见的将来，可视分析这一学科在我国将有巨大的需求和飞速的发展。我国在科技和经济各方面的飞速发展，迫切需要可视化研究提供可视分析数据和诠释的工具。

包括“天河”在内的我国的超级计算机的运算能力已经在国际上处于前列。而运算能力提高产生的巨大的数据；在大飞机等复杂系统开展的研究开发工作，涉及复杂湍流系统等的计算和可视分析；地质勘探油气资源和理解地震机理等地球物理相关研究，也需要大规模数据的可视化。同时，随着互联网的发展，例如微博等服务产生了大量复杂的数据；物联网的发展，对可视分析也有巨大的需求。在国防、国家安全方面，信息化的进步对可视化也提出了相应的要求。

在企业方面，业务数据的增长，一些规模较大的企业将对可视化的研究和应用有强烈的需求。在华跨国公司研究机构和国内的民族企业都加速在可视化方向的研究布局。

IBM 中国研究院、微软亚洲研究院等跨国公司在华研究机构都对可视化特别是信息可视化和可视分析方向给予了高度重视。微软、IBM、微策略、惠普等国际大公司已经设立专门的部门开展面向业务应用的可视化研发。部分互联网企业、通信企业、电力企业等也相继开拓相关方面的工作。国内的大型 IT 企业（如百度公司、阿里集团、腾讯公司等）都相继成立了数据可视化部门或小组。

4.4 三维打印

随着三维打印技术的成熟与普及，三维打印中的计算机图形学研究引起了研究者的极大重视，逐渐成为国内外研究的热点之一。研究工作主要集中在几何优化、结构分析、材料表面效果定制、机构设计和特殊结构设计等方面，这些工作的最终研究目的都是为了使三维模型具有更好的可打印性，能更好地服务于三维打印需要。

在几何优化工作上，国内外学者对分割都做了一些研究，如国外普林斯顿大学的 Chopper 系统，国内中国矿业大学 Hao Jingbin 等的基于曲率的模型分割方法。而国内学者对于打印成本优化问题则给予了更多的关注，如中国科技大学的刘利刚教授团队和山东大学的陈宝权团队分别对此给出了不同方案实现结果。结构分析方面，国外学者如普度大学的 Stava 提出的 Stress Relief 方案，纽约大学的 Qingnan Zhou 给出的 Worst Case 结构分析方法。国内浙江大学周昆团队则对逆向弹性形状设计问题做了一些很好的研究。

在材料表面效果定制工作上，国外研究机构和学者做出了大量突出成果，主要集中在哈佛大学、麻省理工学院、迪斯尼研究中心、瑞士苏黎世联邦理工学院、Adobe 公司、伦敦大学学院这些研究机构。而国内主要是清华大学和微软亚洲研究院的学者如 Dong Yue 等在这方面做了一些很好的工作。

机构设计工作方面，国外研究工作主要集中在积块式机构、动态玩具机构和关节机构方面，完成机构包括新加坡南洋理工大学、迪斯尼苏黎世研究中心、瑞士洛桑联邦理工学院、哈佛大学、伦敦大学学院等。国内的工作主要集中在玩具机构、免组装机构设计方面，这部分工作主要由杭州师范大学和华南理工大学等高校完成。在自支撑结构设计工作上，国外机构如哥伦比亚大学、加州理工学院、瑞士苏黎世联邦理工学院、新加坡南洋理工大学等做了很多工作，国内主要是微软亚洲研究院的 Liu Yan 做了一些研究。内部结构设计工作主要是由国外机构如迪斯尼苏黎世研究中心、卡内基梅隆大学等完成。

5 发展趋势及展望

5.1 图形绘制

三维游戏、三维电影等娱乐产业的飞速发展，三维快速建模、三维获取手段的大力

发展，以及智慧城市建设、大数据可视化等大规模战略需求，为我国图形学的发展带来了新的机遇和挑战。在图形绘制领域，如何快速、真实地呈现复杂三维模型是十分重要的课题。在如下几个方面的研究已经或将成为图形绘制研究的热点：

首先，针对越来越容易获取、数据量越来越大的三维模型数据、物体表面材质数据，图形绘制算法需要向大数据迁移，这需要研究针对海量数据规模的图形处理方法、绘制方法、加速方法以及针对大数据应用的新的绘制硬件、硬件绘制架构等内容。

其次，在计算机相关领域，例如机器学习、数据挖掘等方面，近年来涌现出一些新的分析与处理方法，例如卷积神经网络、矩阵低秩性分析等，这些分析方法已经或可以被应用到传统绘制算法的改进上。通过这些改进，可以进一步提高传统绘制算法的应用范围，并提高绘制的速度。

再次，三维游戏、移动计算平台与高分辨率电视的发展，提高了实时绘制算法的需求，而硬件处理技术的进步则提供了算法改进的硬件平台，这将进一步促进传统真实感绘制算法向实时绘制算法的转化，并进一步发展出新的实时绘制算法。

5.2 数字媒体

近年来，数字媒体的表现形式逐渐向高维发展。从最初的文字、声音、图像、视频，发展到深度图、双目立体影像，到现在的光场、多视点图像、直至真三维图像，高维度形式能够表达更丰富的场景信息，因此必将成为未来的发展趋势。数字媒体表现形式的改变，将催生一整套相应获取、显示、传输和处理方法的更新换代。

人类认知的可计算模型将是数字媒体智能处理的突破口，基于互联网资源的媒体信息交互与合成是构建高质量可视媒体的有效途径。

5.3 可视化与可视分析

随着数据挖掘和可视化研究的逐渐融合，可视分析学的研究目标逐渐从数据规律的挖掘延伸到推理和决策。如何结合相关学科和应用领域，开发高度集成的可视分析系统是未来一个重大的研究课题。

从具体研究方向来看，科学数据、文本数据、高维时空数据和社会网络数据依然是研究重点。科学数据可视化将更多地借鉴信息可视化的交互方式和可视分析的迭代分析流程；信息可视化将逐渐完善可视化的设计空间理论和可视化的评估方法。可视分析将深入更多其他领域，通过巧妙地设计人工干预，将领域专家的智慧加入到分析任务中去，这种人工干预需要考虑人的认知行为特征，通过交互和界面引导用户的可视推理行为。此外，可视分析的任务不再局限于一两种数据类型，将包括实际生产生活环境中的各种类型数据的案例，这种趋势从 2013 年的可视分析挑战赛[⊖]的实时电影票房预测和大规模

[⊖] <http://vacommunity.org/VAST+Challenge+2013>.

网络监控数据（包括文本信息、时序信息、关系信息）分析两个任务可见一斑。

Tableau 的成功上市必将激发工业界对可视化和可视分析的热情，可视分析的工业化将在商业智能外的其他领域大展拳脚，如以性能可视监控见长的 Splunk[⊖]。可视化尤其是可视分析，是一个非常应用化的方向，很多项目工作已经达到了系统原型甚至更成熟，Tableau 本身也源自于斯坦福的项目。工业界应该多关注可视化的工作成果，寻找下一个商机。

5.3 三维打印

目前，三维打印的计算机图形学研究仍处于发展阶段，存在大量有待解决的问题，也是未来方面的研究重点和可能的发展方向，下面试分别略作介绍。

（1）高效便捷的三维建模方法

如前所述，三维模型是三维打印的对象与内容。它是三维打印的信息来源，没有它，三维打印就成了无源之水、无本之木。因此，对三维打印来说，如何能让普通用户高效、便捷地获取生成所需要的三维模型，就是一个需首要解决的任务。因此，如果能为普通用户提供一个便捷、高效的建模技术与工具，必将大大推动三维打印的普及与应用。这方面已经有一些基于草图或笔划的建模方法相关研究，效果还不错，如清华大学和以色列特拉维夫大学的研究人员最近开发出一种名为“3-Sweep”的技术^[124]，可以实现从单张二维照片直接生成三维模型，让三维建模变得像在 Photoshop 中建立选区、编辑图像一样简单。

（2）结构拓扑优化研究

如何通过几何计算方法来优化三维模型的拓扑结构，使其既能外形上满足需要同时结构功能上也有一个很好的性能。这样，既可以减少打印耗材用量，也能缩短打印时间，同时还减少了设备损耗，达到一举三得的良好效果。近来，关于结构拓扑优化已有多篇文章，如 Wang 等^[125~127]，Allaire 等^[128,129]。但如何将这些研究成果运用到三维打印中，仍需要结合三维打印实际需求来考虑，其中还存在很多问题尚待解决。

（3）快速打印研究

三维打印一个合适大小的物体需要一定的耗时，且一般动辄就是十几小时或几十小时，如通过 FDM 工艺采用普通精度打印一个 $40 \times 30 \times 80$ 厘米大小的人头模型约需 12 小时，如果采用更高精度，时间还会更长。这无论是对于个人还是企业，时间成本都有些高，尤其是对企业来说，会大大削弱企业的产品竞争力。因此，缩短三维打印的时间，实现产品对象的快速打印，亟需解决。当然，要想大幅度地缩短打印时间，必须从硬件出发去考虑，改变打印工艺方法，优化打印流程，才能得到较好效果。但是这种方式可能需要付出的代价也非常高。

另一种可能的选择是可以从模型出发，通过几何计算方法来实现快速打印。由于对于三维打印来说，打印时间与打印精度成正比，即打印精度越高，所需打印时间越长。

[⊖] <http://www.splunk.com/>.

因此，如果想要缩短打印时间，一种可行的方法是将模型上一些不重要的部分用低精度，而重要部位用高精度，这样既可缩短打印时间，又不致过于影响模型外观视觉效果，这方面尚存在很多问题值得我们深入探讨和研究。

6 结束语

计算机辅助设计与图形学是与制造业、文化产业、国家安全等密切相关的研究领域，我国学者在基础理论与方法方面取得了一批具有国际影响力的研究成果，在面向领域的应用方面也取得瞩目的成绩。目前，在以三维打印、大数据为代表的需求驱动下，科研人员将和产业界密切合作，为计算机辅助设计与图形学学科发展注入新的活力，带来新的机遇和方向，有力促进与其他学科的交叉融合。同时，我国的研究人员还将为发展我国具有完全自主知识产权的核心基础软件而不懈努力。

致谢

衷心感谢参与本报告撰写、编辑与审定的人员！

参与本研究进展编写的人员包括（按拼音排序）：陈为教授（浙江大学，可视化与可视分析）、刘利刚教授（中国科学技术大学，三维打印）、王锐副教授（浙江大学，图形绘制）、张松海副教授（清华大学，数字媒体）。

浙江大学鲍虎军教授、冯结青教授编辑并审定了全文。博士生陈雪参与了本文的排版与编辑工作。

参考文献

- [1] James Manyika, Michael Chui. Big data: The next frontier for innovation, competition, productivity [R]. McKinsey Global Institute, 2011.
- [2] Travis Korte. Frontiers in Massive Data Analysis. Committee on the Analysis of Massive Data, national-research-council, 2014.
- [3] Benjamin Mora. 2011. Naive ray-tracing: A divide-and-conquer approach [J]. ACM Trans. Graph. 30, 5, Article 117 (October 2011), 12 pages.
- [4] Wenzel Jakob and Steve Marschner. 2012. Manifold exploration: a Markov Chain Monte Carlo technique for rendering scenes with difficult specular transport [J]. ACM Trans. Graph. 31, 4, Article 58 (July 2012), 13 pages.
- [5] Jaakko Lehtinen, Tero Karras, Samuli Laine, Miika Aittala, Frédo Durand, Timo Aila. 2013. Gradient-domain metropolis light transport [J]. ACM Trans. Graph. 32, 4, Article 95 (July 2013), 12 pages.

- [6] Jaakko Lehtinen, Tero Karras, Samuli Laine, Miika Aittala, Frédo Durand, Timo Aila. 2013. Gradient-domain metropolis light transport[J]. ACM Trans. Graph. 32, 4, Article 95 (July 2013), 12 pages.
- [7] Toshiya Hachisuka and Henrik Wann Jensen. 2011. Robust adaptive photon tracing using photon path visibility[J]. ACM Trans. Graph. 30, 5, Article 114 (October 2011), 11 pages.
- [8] Toshiya Hachisuka, Jacopo Pantaleoni, Henrik Wann Jensen. 2012. A path space extension for robust light transport simulation[J]. ACM Trans. Graph. 31, 6, Article 191 (November 2012), 10 pages.
- [9] Iliyan Georgiev, Jaroslav Kivánek, Tomáš Davidovi, Philipp Slusallek. 2012. Light transport simulation with vertex connection and merging[J]. ACM Trans. Graph. 31, 6, Article 192 (November 2012), 10 pages.
- [10] Bruce Walter, Pramook Khungurn, Kavita Bala. 2012. Bidirectional lightcuts[J]. ACM Trans. Graph. 31, 4, Article 59 (July 2012), 11 pages.
- [11] Rui Wang, Yuchi Huo, Yazhen Yuan, Kun Zhou, Wei Hua, Hujun Bao. 2013. GPU-based out-of-core many-lights rendering[J]. ACM Trans. Graph. 32, 6, Article 210 (November 2013), 10 pages.
- [12] Wojciech Jarosz, Volker Schönenfeld, Leif Kobbelt, Henrik Wann Jensen. 2012. Theory, analysis and applications of 2D global illumination[J]. ACM Trans. Graph. 31, 5, Article 125 (September 2012), 21 pages.
- [13] Jaakko Lehtinen, Timo Aila, Samuli Laine, Frédo Durand. 2012. Reconstructing the indirect light field for global illumination[J]. ACM Trans. Graph. 31, 4, Article 51 (July 2012), 10 pages.
- [14] Soham Uday Mehta, Brandon Wang, Ravi Ramamoorthi. 2012. Axis-aligned filtering for interactive sampled soft shadows[J]. ACM Trans. Graph. 31, 6, Article 163 (November 2012), 10 pages.
- [15] Laurent Belcour, Cyril Soler, Kartic Subr, Nicolas Holzschuch, Fredo Durand. 2013. 5D Covariance tracing for efficient defocus and motion blur[J]. ACM Trans. Graph. 32, 3, Article 31 (July 2013), 18 pages.
- [16] Fabrice Rousselle, Claude Knaus, Matthias Zwicker. 2012. Adaptive rendering with non-local means filtering[J]. ACM Trans. Graph. 31, 6, Article 195 (November 2012), 11 pages.
- [17] Tzu-Mao Li, Yu-Ting Wu, Yung-Yu Chuang. 2012. SURE-based optimization for adaptive sampling and reconstruction[J]. ACM Trans. Graph. 31, 6, Article 194 (November 2012), 9 pages.
- [18] Tom Cuypers, Tom Haber, Philippe Bekaert, Se Baek Oh, Ramesh Raskar. 2012. Reflectance model for diffraction[J]. ACM Trans. Graph. 31, 5, Article 122 (September 2012), 11 pages.
- [19] Kun Xu, Wei-Lun Sun, Zhao Dong, Dan-Yong Zhao, Run-Dong Wu, Shi-Min Hu, Anisotropic Spherical Gaussians[J]. ACM Transactions on Graphics 32(6), 209; 1-209; 11, 2013. (Proceedings of SIGGRAPH Asia 2013).
- [20] Roland Ruiters, Christopher Schwartz, Reinhard Klein, Data Driven Surface Reflectance from Sparse and Irregular Samples[J]. Computer Graphics Forum, Volume 31, Issue 2pt1, pages 315-324, May 2012.
- [21] Ioannis Gkioulekas, Bei Xiao, Shuang Zhao, Edward H Adelson, Todd Zickler, Kavita Bala. 2013. Understanding the role of phase function in translucent appearance[J]. ACM Trans. Graph. 32, 5, Article 147 (October 2013), 19 pages.
- [22] Rui Wang, Minghao Pan, Weifeng Chen, Zhong Ren, Kun Zhou, Wei Hua, Hujun Bao, Analytic Double Product Integrals for All-Frequency Relighting [J]. IEEE Transactions on Visualization and Computer Graphics (TVCG), vol. 19, no. 7, pp. 1133-1142, July 2013.
- [23] Kei Iwasaki, Yoshinori Dobashi, Tomoyuki Nishita. 2012. Interactive bi-scale editing of highly glossy materials[J]. ACM Trans. Graph. 31, 6, Article 144 (November 2012), 7 pages.
- [24] Jan Novák, Derek Nowrouzezahrai, Carsten Dachsbacher, Wojciech Jarosz. 2012. Virtual ray lights for

- rendering scenes with participating media[J]. ACM Trans. Graph. 31, 4, Article 60 (July 2012), 11 pages.
- [25] Ioannis Gkioulekas, Shuang Zhao, Kavita Bala, Todd Zickler, Anat Levin. 2013. Inverse volume rendering with material dictionaries[J]. ACM Trans. Graph. 32, 6, Article 162 (November 2013), 13 pages.
- [26] Petrik Clarberg, Robert Toth, Jacob Munkberg. 2013. A sort-based deferred shading architecture for decoupled sampling[J]. ACM Trans. Graph. 32, 4, Article 141 (July 2013), 10 pages.
- [27] Rasmus Barringer, Tomas Akenine-Möller. 2013. A4: asynchronous adaptive anti-aliasing using shared memory[J]. ACM Trans. Graph. 32, 4, Article 100 (July 2013), 10 pages.
- [28] Michael J Doyle, Colin Fowler, Michael Manzke. 2013. A hardware unit for fast SAH-optimised BVH construction[J]. ACM Trans. Graph. 32, 4, Article 139 (July 2013), 10 pages.
- [29] Niesner and Loop, 2013] Matthias Nießner and Charles Loop. 2013. Analytic displacement mapping using hardware tessellation[J]. ACM Trans. Graph. 32, 3, Article 26 (July 2013), 9 pages.
- [30] Jiating Chen, Xiaoyin Ge, Li-Yi Wei, Bin Wang, Yusu Wang, Huamin Wang, Yun Fei, Kang-Lai Qian, Jun-Hai Yong, Wenping Wang. 2013. Bilateral blue noise sampling[J]. ACM Trans. Graph. 32, 6, Article 216 (November 2013), 11 pages.
- [31] Xin Sun, Kun Zhou, Jie Guo, Guofu Xie, Jingui Pan, Wencheng Wang, Baining Guo. 2013. Line segment sampling with blue-noise properties[J]. ACM Trans. Graph. 32, 4, Article 127 (July 2013), 14 pages.
- [32] Ravi Ramamoorthi, John Anderson, Mark Meyer, Derek Nowrouzezahrai. 2012. A theory of monte carlo visibility sampling[J]. ACM Trans. Graph. 31, 5, Article 121 (September 2012), 16 pages.
- [33] Lin S, Ritchie D, Fisher M, Hanrahan P. Probabilistic Color-by-Numbers: Suggesting Pattern Colorizations Using Factor Graphs[J]. Acm Transactions on Graphics, 2013, 32(4). doi: Artn 37.
- [34] Limpachet, A, Feltman, N, Treuille, A, & Cohen, M. Real-time Drawing Assistance through Crowdsourcing [J]. Acm Transactions on Graphics, 2013, 32(4). doi: Artn 54.
- [35] Berger, I, Shamir, A, Mahler, M, Carter, E, & Hodgins, J. Style and Abstraction in Portrait Sketching [J]. Acm Transactions on Graphics, 2013, 32(4). doi: Artn 55.
- [36] Bouaziz et al., 2013] Bouaziz S, Wang Y G, Pauly M. Online Modeling For Realtime Facial Animation[J]. Acm Transactions on Graphics, 2013, 32(4). doi: Artn 40.
- [37] Cao C, Weng Y L, Lin S, Zhou K. 3D Shape Regression for Real-time Facial Animation [J]. Acm Transactions on Graphics, 2013, 32(4). doi: Artn 41.
- [38] Wang Y G, Min J Y, Zhang J J, Liu Y B, Xu F, Dai Q H, Chai J X. Video-based Hand Manipulation Capture Through Composite Motion Control [J]. Acm Transactions on Graphics, 2013, 32 (4). doi: Artn 43.
- [39] Ali-Hamadi D, Liu T T, Gilles B, Kavan L, Faure F, Palombi O, Cani M P. Anatomy Transfer[J]. Acm Transactions on Graphics, 2013, 32(6). doi: Artn 188.
- [40] Guay M, Cani M P, Ronfard R. The Line of Action: an Intuitive Interface for Expressive Character Posing [J]. Acm Transactions on Graphics, 2013, 32(6). doi: Artn 205.
- [41] Hoyet L, Ryall K, Zibrek K, Park H, Lee J, Hodgins J, O'Sullivan C. Evaluating the Distinctiveness and Attractiveness of Human Motions on Realistic Virtual Bodies[J]. Acm Transactions on Graphics, 2013, 32 (6). doi: Artn 204.
- [42] Tony Hey. 第四范式：数据密集型科学发现[M]. 科学出版社, 2012.
- [43] Ayan Biswas, Soumya Dutta, Han-Wei Shen, Jonathan Woodring, An Information-Aware Framework for Exploring Multivariate Data Sets[J]. IEEE Transactions on Visualization and Computer Graphics, 2013.

- [44] Fang Zheng, Hongfeng Yu, Can Hantas, Matthew Wolf, Greg Eisenhauer, Karsten Schwan, Hasan Abbasi, Scott Klasky. GoldRush: Resource Efficient In Situ Scientific Data Analytics Using Fine-Grained Interference Aware Execution[C]. ACM/IEEE Supercomputing Conference (SC), November, 2013.
- [45] M Chen, D Ebert, H Hagen, R S Laramee, R van Liere, K-L Ma, W Ribarsky, G Scheuermann, D Silver. Data, Information and Knowledge in Visualization[J], IEEE Computer Graphics and Applications, 2009, 29(1) : 12-19.
- [46] Benjamin Bach, Emmanuel Pietriga, Jean-Daniel Fekete. GraphDiaries: Animated Transitions and Temporal Navigation for Dynamic Networks [J]. IEEE Transactions on Visualization and Computer Graphics (TVCG), 2014, 20 (5) : 740 - 754.
- [47] Muhlbacher T, Piringer H. A partition-based framework for building and validating regression models[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 1962-1971.
- [48] Bogl M, Aigner W, Filzmoser P, et al. Visual analytics for model selection in time series analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2237-2246.
- [49] Gleicher M Explainers: Expert Explorations with Crafted Projections[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2042-2051.
- [50] Schmidt J, Groller M E, Bruckner S VAICo; Visual Analysis for Image Comparison[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2090-2099.
- [51] Meghdadi A H, Irani P. Interactive Exploration of Surveillance Video through Action Shot Summarization and Trajectory Visualization [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19 (12) : 2119-2128.
- [52] Schultz T, Kindlmann G L Open-Box Spectral Clustering: Applications to Medical Image Analysis[J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2100-2108.
- [53] Brehmer M, Munzner T. A multi-level typology of abstract visualization tasks [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2376-2385.
- [54] Luo Linjie, Baran Ilya, Rusinkiewicz Szymon, Matusik Wojciech. Chopper: partitioning models into 3D-printable parts[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2012), 2012, 31(6) : 129 ; 1-129 : 10.
- [55] Chen Desai, Sithi-amorn Pitchaya, Lan Justin T, Matusik Wojciech. Computing and Fabricating Multiplanar Models[J]. Computer Graphics Forum, 2013, 32(2pt3) : 305-315.
- [56] Prévost Romain, Whiting Emily, Lefebvre Sylvain, Sorkine-Hornung Olga. Make It Stand: Balancing Shapes for 3D Fabrication[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32 (4) : 81 : 1-81 : 10.
- [57] Stava Ondrej, Vanek Juraj, Benes Bedrich, Carr Nathan, Měch Radomír. Stress relief: improving structural strength of 3D printable objects [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2012), 2012, 31(4) : 48 : 1-48 : 11.
- [58] Zhou Qingnan, Panetta Julian, Zorin Denis. Worst-Case Structural Analysis [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 137 ; 1-137 : 12.
- [59] Hašan Miloš, Fuchs Martin, Matusik Wojciech, Pfister Hanspeter, Rusinkiewicz Szymon. Physical reproduction of materials with specified subsurface scattering [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2010), 2010, 29(4) : 61 : 1-61 : 10.
- [60] Papas Marios, Regg Christian, Jarosz Wojciech, Bickel Bernd, Jackson Philip, Matusik Wojciech,

- Marschner Steve, Gross Markus. Fabricating translucent materials using continuous pigment mixtures [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4): 146: 1-146: 12.
- [61] Weyrich Tim, Peers Pieter, Matusik Wojciech, Rusinkiewicz Szymon. Fabricating microgeometry for custom surface reflectance[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2009), 2009, 28(3): 32: 1-32: 6.
- [62] Matusik Wojciech, Ajdin Boris, Gu Jinwei, Lawrence Jason, Lensch Hendrik, Pellacini Fabio, Rusinkiewicz Szymon. Printing spatially-varying reflectance [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2009), 2009, 28(5): 128: 1-128: 10.
- [63] Malzbender Tom, Samadani Ramin, Scher Steven, Crume Adam, Dunn Douglas, Davis James. Printing reflectance functions[J]. ACM Transactions on Graphics, 2012, 31(3): 20: 1-20: 11.
- [64] Levin Anat, Glasner Daniel, Xiong Ying, Durand Frédéric, Freeman William, Matusik Wojciech, Zickler Todd. Fabricating BRDFs at High Spatial Resolution Using Wave Optics[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4): 144: 1-144: 13.
- [65] Bickel Bernd, Bächer Moritz, Otaduy Miguel A, Lee Hyunho Richard, Pfister Hanspeter, Gross Markus, Matusik Wojciech. Design and fabrication of materials with desired deformation behavior [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2010), 2010, 29(4): 63: 1-63: 10.
- [66] Chen Desai, Levin David IW, Didyk Piotr, Sitthi-Amorn Pitchaya, Matusik Wojciech. Spec2Fab: a reducer-tuner model for translating specifications to 3D prints [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4): 135: 1-135: 10.
- [67] Xin ShiQing, Lai Chifu, Fu Chiwing, Wong Tientsin, He Ying, Cohen-Or Daniel. Making burr puzzles from 3D models[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2011), 2011, 30(4): 97: 1-97: 8.
- [68] Song Peng, Fu Chiwing, Cohen-Or Daniel. Recursive interlocking puzzles [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2012), 2012, 31(6): 128: 1-128: 10.
- [69] Schwartzburg Yuliy, Pauly Mark. Fabrication - aware Design with Intersecting Planar Pieces[J]. Computer Graphics Forum, 2013, 32(2pt3): 317-326.
- [70] Coros Stelian, Thomaszewski Bernhard, Noris Gioachino, Sueda Shinjiro, Forberg Moira, Sumner Robert W, Matusik Wojciech, Bickel Bernd. Computational design of mechanical characters[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4): 83: 1-83: 12.
- [71] Duygu Ceylan Wilmot Li, Niloy J Mitra, Maneesh Agrawala, Mark Pauly. Designing and Fabricating Mechanical Automata from Mocap Sequences[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2013), 2013, 32(6): 186: 1-186: 11.
- [72] Calì Jacques, Calian Dan A, Amati Cristina, Kleinberger Rebecca, Steed Anthony, Kautz Jan, Weyrich Tim. 3D-printing of non-assembly, articulated models[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2012), 2012, 31(6): 130: 1-130: 8.
- [73] Bächer Moritz, Bickel Bernd, James Doug L, Pfister Hanspeter. Fabricating articulated characters from skinned meshes[J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2012), 2012, 31(4): 47: 1-47: 9.
- [74] Block Philippe, Ochsendorf John. Thrust Network Analysis: A new methodology for three-dimensional equilibrium[J]. Journal of the International Association for Shell and Spatial Structures, 2007, 155(3): 167-174.

- [75] Panizzo Daniele, Block Philippe, Sorkine-Hornung Olga. Designing unreinforced masonry models [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 91: 1-91; 12.
- [76] De Goes Fernando, Alliez Pierre, Owhadi Houman, Desbrun Mathieu. On the Equilibrium of Simplicial Masonry Structures [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 93: 1-93; 10.
- [77] Vouga Etienne, Höbinger Mathias, Wallner Johannes, Pottmann Helmut. Design of self-supporting surfaces [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2012), 2012, 31 (4) : 87: 1-87: 11.
- [78] Song Peng, Fu Chiwing, Goswami Prashant, Zheng Jianmin, Mitra Niloy J, Cohen-Or Daniel. Reciprocal frame structures made easy [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 94: 1-94; 10.
- [79] Willis Karl DD, Wilson Andrew D. InfraStructs: fabricating information inside physical objects for imaging in the terahertz region [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 138: 1-138; 10.
- [80] Holroyd Michael, Baran Ilya, Lawrence Jason, Matusik Wojciech. Computing and fabricating multilayer models [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2011), 2011, 30(6) : 187: 1-187; 8.
- [81] Peiran Ren, Jiaping Wang, Minnin Gong, Stephen Lin, Xin Tong, Baining Guo, Global Illumination with Radiance Regression Functions [J]. ACM Transaction on Graphics, Vol. 32, No. 4, Article 130, 2013 (SIGGRAPH 2013).
- [82] Rui Wang, Yingqing Wu, Minghao Pan, Wei Chen, Wei Hua, Shadow geometry maps for alias-free shadows [J]. SCIENCE CHINA Information Sciences, 2012, 55(11).
- [83] Li Shen, Jieqing Feng, Baoguang Yang, Exponential Soft Shadow Mapping [J]. Computer Graphics Forum (Special issue of EGSR2013), 2013, 32(4) : 107-116.
- [84] Rui Wang, Minghao Pan, Xiang Han, Weifeng Chen, Hujun Bao, Parallel and Adaptive Visibility Sampling for Rendering Dynamic Scenes with Spatially-Varying Reflectance [J]. To appear in Computer & Graphics special issue on CAD/GRAFPHICS 2013.
- [85] Ling-Qi Yan, Yahan Zhou, Kun Xu, Rui Wang, Accurate Translucent Material Rendering under Spherical Gaussian Lights [J]. Computer Graphics Forum (Proceedings of Pacific Graphics 2012), 2012, 31 (7) : 2267-2276.
- [86] Dongping Li, Xin Sun, Zhong Ren, Steve Lin, Yiyi Tong, Baining Guo, Kun Zhou, TransCut: Interactive Rendering of Translucent Cutouts [J]. IEEE Transactions on Visualization & Computer Graphics, 2013.
- [87] Hao Qin, Menglei Chai, Qiming Hou, Zhong Ren, Kun Zhou, Cone Tracing for Furry Object Rendering [J]. IEEE TVCG, 2014.
- [88] Zhonggui Chen, Zhan Yuan, Yi-King Choi, Ligang Liu, Wenping Wang. Variational Blue Noise Sampling [J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18: 1784-1796.
- [89] Ye G Z, Liu Y B, Deng Y, Hasler N, Ji X Y, Dai Q H, Theobalt C. Free-Viewpoint Video of Human Actors Using Multiple Handheld Kinects [J]. IEEE Transactions on Cybernetics, 2013, 43 (5) : 1370-1382.
- [90] Fu Q, Liu Y-J, Chen W, Fu X. The time course of natural scene categorization in human brain: simple line-

- drawings vs. color photographs [J]. *Journal of Vision*, 2013, 13(9) : 1060.
- [91] Yong-Jin L, Xi L, Joneja A, Cui-Xia M, Xiao-Lan F, Dawei S. User-Adaptive Sketch-Based 3-D CAD Model Retrieval [J]. *IEEE Transactions on Automation Science and Engineering*, 2013, 10(3) : 783-795.
- [92] Hu S M, Zhang F L, Wang M, Martin R R, Wang J. PatchNet: A Patch-based Image Representation for Interactive Library-driven Image Editing [J]. *Acm Transactions on Graphics*, 2013, 32(6). doi: Artn 196.
- [93] Xu K, Chen K, Fu H B, Sun W L, Hu S M. Sketch2Scene: Sketch-based Co-retrieval and Co-placement of 3D Models [J]. *Acm Transactions on Graphics*, 2013, 32(4). doi: Artn 123.
- [94] Chen T, Zhu Z, Shamir A, Hu S M, Cohen-Or D. 3-Sweep: Extracting Editable Objects from a Single Photo [J]. *Acm Transactions on Graphics*, 2013, 32(6). doi: Artn 195.
- [95] Du S P, Masia B, Hu S M, Gutierrez D. A Metric of Visual Comfort for Stereoscopic Motion [J]. *Acm Transactions on Graphics*, 2013, 32(6). doi: Artn 222.
- [96] Du S P, Hu S M, Martin R R. Changing Perspective in Stereoscopic Images [J]. *IEEE Transactions on Visualization And Computer Graphics*, 2013, 19(8) : 1288-1297.
- [97] Li X Y, Gu Y, Hu S M, Martin R R. Mixed-Domain Edge-Aware Image Manipulation [J]. *IEEE Transactions on Image Processing*, 2013, 22(5) : 1915-1925.
- [98] Huang S S, Shamir A, Shen C H, Zhang H, Sheffer A, Hu S M, Cohen-Or D. Qualitative Organization of Collections of Shapes via Quartet Analysis [J]. *Acm Transactions on Graphics*, 2013, 32(4). doi: Artn 71.
- [99] Hu S M, Xu K, Ma L Q, Liu B, Jiang B Y, Wang J. Inverse Image Editing: Recovering a Semantic Editing History from a Before-and-After Image Pair [J]. *Acm Transactions on Graphics*, 2013, 32 (6). doi: Artn 194.
- [100] Guo H, Yuan X. Local WYSIWYG volume visualization [C], *IEEE Pacific Visualization Symposium (PacificVis)*, 2013 : 65-72.
- [101] Yang J, Liu Y, Zhang X, et al. PIWI: Visually exploring graphs based on their community structure [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(6) : 1034-1047.
- [102] Yuan X, Ren D, Wang Z, et al. Dimension Projection Matrix/Tree: Interactive Subspace Visual Exploration and Analysis of High Dimensional Data [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12) : 2625-2633.
- [103] Wang Z, Lu M, Yuan X, et al. Visual traffic jam analysis based on trajectory data [J]. *IEEE Transactions on Visualization and Computer Graphics*, 2013, 19(12) : 2159-2168.
- [104] GuiZhen Wang, Chaokai Wen, BingHui Yan, Cong Xie, Ronghua Liang, Wei Chen. Topic Hypergraph: Hierarchical Visualization of Thematic Structures in Long Documents Science in China. 2013, Vol. 56 052111 : 1-052111 : 14.
- [105] Yuxin Ma, Jiayi Xu, Dichao Peng, Ting Zhang, Chengzhe Jin, Huamin Qu, Wei Chen, Qunsheng Peng. A Visual Analysis Approach for Community Detection of Multi-Context Mobile Social Networks [J]. *Journal of Computer Science and Technology*, 2013, 28(5) : 797-809.
- [106] Jing Xia, Feiran Wu, Fangzhou Guo, Cong Xie, Zhen Liu, Wei Chen. An Online Visualization System for Streaming Log Data of Computing Clusters [J]. 清华大学学报(英文版), 2013, 18(2) : 196-205.
- [107] Qiang Zhao, Liang Wan, Wei Feng, Tien-Tsin Wong, Jiawan Zhang: Cube2Video: Navigate between Cubic Panoramas in Real-Time [J]. *IEEE Transactions on Multimedia*, 2013.
- [108] Nguyen Quang, Vinh, Qian Yu, Huang MaoLin & Zhang JiaWan: TabuVis: A tool for visual analytics multidimensional datasets [J]. *SCIENCE CHINA Information Sciences*, 2013, 56.

- [109] Zhang J, Kang K, Liu D, et al. Vis4Heritage: Visual Analytics Approach on Grotto Wall Painting Degradations [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19 (12) : 1982-1991.
- [110] Xu P, Du F, Cao N, et al. Visual Analysis of Set Relations in a Graph [J]. Computer Graphics Forum. Blackwell Publishing Ltd, 2013, 32(3pt1) : 61-70.
- [111] Zeng W, Fu C W, Arisona S M, et al. Visualizing Interchange Patterns in Massive Movement Data [J]. Computer Graphics Forum. Blackwell Publishing Ltd, 2013, 32(3pt3) : 271-280.
- [112] Xu P, Wu Y, Wei E, et al. Visual analysis of topic competition on social media [J]. IEEE Transactions on Visualization and Computer Graphics, 2013, 19(12) : 2012-2021.
- [113] Hao Jingbin, Fang Liang, Williams Robert E. An efficient curvature-based partitioning of large-scale STL models [J]. Rapid Prototyping Journal, 2011, 17(2) : 116-127.
- [114] Wang Weiming, Wang Tuanfeng Y, Yang Zhouwang, Liu Ligang, Tong Xin, Tong Weihua, Deng Jiansong, Chen Falai, Liu Xiuping. Cost-effective Printing of 3D Objects with Skin-Frame Structures [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2013), 2013, 32(6) : 177: 1-177: 10.
- [115] Lu Lin, Sharf Andrei, Zhao Haisen, Wei Yuan, Fan Qingnan, Chen Xuelin, Savoye Yann, Tu Changhe, Cohen-Or Daniel, Chen Baoquan. Build-to-Last: Strength to Weight 3D Printed Objects [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2014), 2014, 33(4).
- [116] Chen Xiang, Zheng Changxi, Xu Weiwei, Zhou Kun. An Asymptotic Numerical Method for Inverse Elastic Shape Design [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2014), 2014, 33(4).
- [117] Dong Yue, Lin Stephen, Guo Baining. Fabricating spatially-varying subsurface scattering [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2010), 2010, 29(4) : 153: 1-153: 10.
- [118] Dong Yue, Tong Xin, Pellacini Fabio, Guo Baining. Printing spatially-varying reflectance for reproducing HDR images [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2012), 2012, 31 (4) : 40: 1-40: 8.
- [119] Lan Yanxiang, Dong Yue, Pellacini Fabio, Tong Xin. Bi-Scale Appearance Fabrication [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 145: 1-145: 12.
- [120] Zhu Lifeng, Xu Weiwei, Snyder John, Liu Yang, Wang Guoping, Guo Baining. Motion-guided mechanical toy modeling [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2012), 2012, 31(6) : 127: 1-127: 10.
- [121] Su Xubin, Yang Yongqiang, Wang Di, Chen Yonghua. Digital assembly and direct fabrication of mechanism based on selective laser melting. Rapid Prototyping Journal, 2013, 19(3) : 166-172.
- [122] Liu Yang, Pan Hao, Snyder John, Wang Wenping, Guo Baining. Computing self-supporting surfaces by regular triangulation [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2013), 2013, 32(4) : 92: 1-92: 10.
- [123] Velten, A., Wu, D., Jarabo, A., Masia, B., Barsi, C., Joshi, C., Raskar, R (2013). Femto-photography: capturing and visualizing the propagation of light [J]. Acm Transactions on Graphics, 32(4).
- [124] Chen Tao, Zhu Zhe, Shamir Ariel, Hu Shi-Min, Cohen-Or Daniel. 3-Sweep: extracting editable objects from a single photo [J]. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH Asia 2013), 2013, 32(6) : 195: 1-195: 10.
- [125] Wang Michael Yu, Wang Xiaoming, Guo Dongming. A level set method for structural topology optimization [J]. Computer methods in applied mechanics and engineering, 2003, 192(1) : 227-246.

- [126] Xing Xianghua, Wei Peng, Wang Michael Yu. A finite element - based level set method for structural optimization [J]. International Journal for Numerical Methods in Engineering, 2010, 82(7) : 805-842.
- [127] Luo Junzhao, Luo Zhen, Chen Liping, Tong Liyong, Wang Michael Yu. A semi-implicit level set method for structural shape and topology optimization [J]. Journal of Computational Physics, 2008, 227 (11) : 5561-5581.
- [128] Allaire Grégoire, Jouve François, Toader Anca-Maria. Structural optimization using sensitivity analysis and a level-set method [J]. Journal of Computational Physics, 2004, 194(1) : 363-393.
- [129] Allaire Grégoire, De Gournay Frédéric, Jouve François, Toader A. Structural optimization using topological and shape sensitivity via a level set method [J]. Control and Cybernetics, 2005, 34(1) : 59-80.

作者简介

鲍虎军 博士, 浙江大学信息学部主任、CAD&CG 国家重点实验室学术委员会副主任、教育部长江学者特聘教授、博士生导师。主要研究方向为计算机图形学、计算机视觉、虚拟现实。CCF 常务理事, 中国计算机学会第七届计算机辅助设计与图形学专业委员会主任。



陈 为 博士, 浙江大学 CAD&CG 国家重点实验室教授、博士生导师。主要研究方向为可视分析、可视化。



冯结青 博士, 浙江大学 CAD&CG 国家重点实验室教授、博士生导师。主要研究方向为计算机图形学。CCF 高级会员, 中国计算机学会第七届计算机辅助设计与图形学专业委员会秘书长。



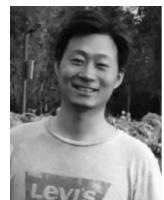
刘利刚 博士, 中国科学技术大学数学科学学院教授、博士生导师。主要研究方向为计算机图形学。中国工业与应用数学学会几何设计与计算专业委员会委员。



王 锐 博士，浙江大学 CAD&CG 国家重点实验室副教授、博士生导师。主要研究方向为计算机图形学、虚拟现实。



张松海 博士，清华大学计算机系副教授。主要研究方向为计算机图形学、图像/视频处理。CCF 高级会员，中国图像图形学学会多媒体专业委员会秘书长。



移动学习的研究进展与趋势

郑庆华¹ 张未展¹ 田 锋² 魏笔凡³ 杜海鹏³

¹西安交通大学电信学院计算机系，西安

²西安交通大学电信学院自动化系，西安

³西安交通大学网络教育学院，西安

摘要

移动学习（m-Learning）作为下一代 e-Learning 系统的典型服务模式，旨在构建具有高度灵活性、移动性、个性化、协作性的网络学习环境，具有重要的研究与应用价值。其研究领域涵盖移动计算、云计算、网络多媒体、知识获取与个性化服务等相关技术，涉及计算机科学、教育学等多门学科。因此，如何真正满足人们的移动化认知学习需求，是一项极富挑战性的研究内容。本报告详细分析并比较了移动学习技术的国内外研究现状，对移动学习技术的发展趋势进行了展望，并特别介绍了 MOOC 与大数据分析对移动学习发展趋势的影响。

关键词：移动学习，微课程，个性化服务，云计算，大数据分析，大规模在线开放课程

Abstract

With the incredible growth of mobile network, Mobile Learning (m-Learning) has became a typical service model of the next generation e-Learning systems, it aims to build a highly flexible, mobile, personalized, and collaborative learning environment, m-Learning shows great value not only in theory but also in application, which needs the support of a series of technologies including mobile computing, cloud computing, multimedia, knowledge acquisition and personalized service, and so on, involving computer science, education and other subjects. This report gives a detailed comparison analysis of state-of-the-art of Mobile Learning technology, and prospects the trends of Mobile Learning. Especially, the impact on Mobile Learning by the big data and the MOOCs is analyzed.

Keywords: Mobile Learning, Microlecture, Personalized service, Cloud computing, Big data, Massive Open Online Courses

1 引言

1.1 移动学习的发展历程

e-Learning 是构建学习型社会和终身学习体系，缓解我国城乡之间、东西部之间日趋

严重的教育数字鸿沟和教育公平问题的基本技术途径，是《国家中长期科学和技术发展规划纲要（2006—2020）》和《国家中长期教育改革和发展规划纲要（2010—2020）》的战略任务之一。随着移动互联网络的迅猛发展，移动学习（Mobile Learning, m-Learning）将成为下一代 e-Learning 系统的主要特征之一。

移动学习有多种定义方式，广义的讲，体现 4A 特性（Anywhere, Anytime, Anyone, Anydevice）的学习方式都可以称作移动学习。据此定义，移动学习者口袋里可放置的静态纸质图书与电子书籍都可以归入移动学习的范畴。本报告所指的移动学习，参考了 Crompton 在 2013 年给出的移动学习最新定义^[1]，即移动学习应同时具备两个特征：1) 使用便携式可移动的个人电子设施；2) 通过社会化的交互方式实现知识的传播与学习。便携式移动电子设备将移动学习从一种学习模式限定在了信息技术领域的范畴，而社会化的交互方式则突出了移动学习的社会化、系统化特征。

广义的移动学习历史可以追溯到 20 世纪 60 年代。早在 1968 年，以富有创新精神著称的施乐帕克研究中心（Xerox PARC）研发出一个电子书式的便携式计算机 Dynabook，用于动态的仿真学习。本报告所侧重的信息化技术与社会化交互相结合的移动学习技术，则起源于 20 世纪 90 年代^[2]。在 20 世纪 90 年代早期，苹果公司推出了“明天课堂”（也译为“未来教室”）。在该项目的“Wireless Coyote”子项目中，基于 Palm 操作系统实现了完整的移动学习系统解决方案。1994 年，美国卡内基·梅隆大学开展了一个名为“WirelessAndrew”的项目。该项目作为学校无线网络基础设施建设的一部分，目的是为全校师生提供一个覆盖校园的无线高速网络，使他们可以在学校的任何地方都能轻松上网学习。

进入 21 世纪，随着移动互联网与移动终端的普及，移动学习发展迅速。在 21 世纪初，欧盟多个国家与组织共同发起 MOBILearn 项目，由欧盟多个国家共同进行内容建设与移动教学服务，并可以提供统一的移动学习用户注册认证。项目针对学习者的需求建立了支持移动学习的 WAP 教育站点，将移动技术和设备应用于在校学习和终身学习。与此同时，在世界范围内，移动学习迅速发展，教育界、学术界、工业界都对移动学习投入了极大的热情。在 2007 年，英国的 MoleNET 项目已经使移动学习从校园走向了厂区，为英国的职业再教育服务。

2010 年以后，随着移动互联网、智能终端、云计算等技术的发展，移动学习进入了一个全新的时代，支持多种移动操作系统的移动学习应用仅仅是移动学习的基础，如何基于泛在性、及时性、情景性、专属性的特征，提供智能化、个性化的移动学习服务，提升移动学习者的学习兴趣与用户体验，提高移动学习的学习效率，是新一代移动学习系统关注的焦点。2011 年，大规模开放在线课程（Massive Open Online Course, MOOC）在全球范围内推广开来，其“非正式、情境性、个性化”的思想与移动学习不谋而合。支持移动学习的 Mobile MOOC，将成为移动学习发展的重要方向。

1.2 移动学习技术研究的重要价值

移动互联网技术的迅猛发展，特别是智能移动设备的普及，推动了传统的 e-Learning

向 m-Learning 的转变。智能移动设备的便携性和移动性，使得人们可以随时随地进行学习，人们的学习方式已经发生改变，移动学习将成为数字化远程教育下一个发展方向，是一种未来学习不可缺少的学习模式。

移动学习技术研究的重要价值与意义体现在两个方面：一方面，移动学习技术的新特征，对提升学习用户的学习方式与学习体验具有重要的价值与意义；另一方面，移动学习应用作为计算机学科多领域技术的载体，对计算机学科相关领域技术的发展产生巨大的推动作用，为计算机学科多个领域带来了新的挑战与机遇。

与数字化远程教育相比，移动学习具有泛在性、及时性。首先，移动学习技术可以使任何人在任何时间、任何地点学习或传播任何知识，任何持有移动终端的人都可以成为移动学习中的学习者和教育者，使得教育得到更广泛的普及。其次，移动学习技术使学习者更方便地学习，也使教育者更便捷地教育。利用具有移动通信功能的智能设备，以及智能设备上开发的专用学习软件，学习者可以及时获取知识，及时进行学习、交流、讨论，可以在需要某些知识的时候随时学习；教育者也可以借助移动网络及时对学生进行辅导。

与数字化远程教育相比，移动学习还具有情景性、专属性。移动学习可以突破数字化远程教育学习过程中需要在电脑或书桌前的限制，可以在移动中利用碎片化时间充分学习。而在碎片化学习中，学习者所处的情境（如位置、周围人群、周围干扰等）变化很快，移动学习技术可自动感知学习者及学习者情境的变化，并推荐符合其所处情境的个性化学习服务，提升学习系统个性化、智能化的程度。

总之，在智能移动设备和移动通信技术迅速发展的背景下，越来越多的学习者使用智能移动终端随时随地进行学习，拓宽了教育的范围，推动了终身教育的进程，对于构建学习型社会意义重大。

另一方面，移动学习的发展也将对各个计算机科学相关技术领域的进步起到推动的作用。

1) 云计算资源调度技术。由于移动学习终端设备的便携性特征，终端的资源将是有有限的，为了提高移动学习的能力，更多的资源将统一在云端管理，移动学习的规模化服务需求，将促进大规模云资源调度技术的发展。

2) 移动网络多媒体传输技术。在移动学习中有大量的音视频教学场景数据需要传输，而移动网络信道具有带宽的不确定性以及数据传输的低可靠性，将影响学生的学习体验。学习者对移动学习体验的质量要求，将促进移动网络多媒体传输技术的发展。

3) 移动终端节能与资源优化利用技术。移动学习终端需要有适合的显示屏和丰富的多媒体编解码程序，这些应用都是电量高消耗型应用，会严重缩短电池的供应时间，给移动学习带来极大的不便。针对移动学习应用能耗大的特点，研究相应的节能策略，在一定程度上可缓解移动学习设备的能耗问题；移动学习的高能耗需求，将促进移动终端节能与资源优化利用技术的发展。

4) 移动视频资源的生成、转换与存储管理技术。由于移动学习终端的功能有限，海量的移动教育资源需要通过云计算技术进行存储、管理、共享，研究短视频移动学习资

源的生成、转换与管理，可最大限度地实现海量学习资源的无缝对接，适应泛在学习的需求。短视频移动学习资源的生成、转换与管理共享需求，可促进移动视频资源的生成、转换与存储管理技术的发展。

5) 大数据分析与个性化服务技术。移动学习是当前碎片化和泛在学习时代获取信息的新型学习模式，具有情境化和个性化的特征。而移动学习系统运用云计算技术存储和管理了大量学习数据，为全面跟踪和掌握学生特点、学习行为、学习过程提供了数据基础，通过大数据分析技术，对个体学习用户进行有针对性、差异化的个性化教学，可更准确地评价学生，提高学生的学习质量和学习效率。移动学习的情境化和个性化特征与应用需求，可促进大数据分析与个性化技术的发展。

2 国际研究现状

本节将对移动学习技术的国际研究现状与面临的挑战进行深入的分析，阐明移动学习技术在信息领域面临的主要问题，介绍已有移动学习技术研究与应用的国际研究现状，并探讨尚未解决的难题与挑战。

下面将分别从移动学习系统架构、移动学习关键技术、移动学习系统部署应用三个角度展开分析与讨论。

2.1 移动学习系统架构的技术演变

移动和无线网络技术的迅速发展推动着国际上关于移动学习系统架构技术的不断演变和进步，从早期的 C/S 或 B/S 架构演变为云端结合的架构。

移动学习发展早期阶段，其功能比较简单，使用的规模也有限，主要采用 C/S 或 B/S 架构，利用移动设备的便捷性、移动性，把学习内容以“推”送方式来传播给学习者，这是一种被动的、单向的、机械的交互方式。这一阶段常见的学习形式有：基于 SMS 或 MMS 的移动学习形式、基于页面浏览的学习形式。这个阶段也采用扩展的 C/S 或 B/S 架构，如全球最大的网络学习方案服务商之一 Blackboard，该公司采用 B/S 架构，应用服务器和数据库服务器完全可以和 Web 服务器剥离，放在局域网内部，防火墙可设置成只对外公开 80 端口（即 HTTP 服务器端口），可有效保证应用和数据的安全性。

随着学习规模的增长，C/S 或 B/S 架构已无法满足移动学习的需求。云计算技术的出现，为解决学习规模的增长提供了一个可行的途径。云计算环境下，所有的数据存储和处理都将在“云”端进行，学习者只需通过浏览器便可进行类似于在个人计算机上的常用操作。“云”是指由基础设施、平台和应用软件构成的综合体系，由大量的并行分布式计算系统服务器组成集群提供原始资源，如硬盘、CPU 和 GPU 等，可用来开发多媒体应用和服务，如存储、编辑、流化、渲染等^[3]。“云”端可以是自己建设的私有云，以方便管理控制。而在国际上，更多使用的是第三方云服务提供商提供的公有云服务，最

新的国外大规模移动学习系统普遍采用 YouTube 等第三方平台提供视频存储、播放、浏览器组件等服务。另外也可以采用亚马逊公司提供的 EC2 虚拟机实现流媒体服务，将视频资源文件存储在云平台中；还可利用 CDN 服务商（如 Akamai）提供的内容分发服务，将视频分散在 CDN 上，用户可以就近访问相应的资源服务器，避免跨网或者远程访问源资源，提高响应速度和服务质量。

2.2 移动学习关键技术的研究现状

移动学习技术并不是一项专有的领域技术，而是以移动学习应用为载体，计算机学科多领域相关技术的实例化与具体化，在移动学习核心关键技术领域，我们将重点从移动计算、云计算、网络多媒体、知识获取与个性化服务等相关技术的角度出发，探讨移动学习技术面临的挑战与机遇。

如图 1 所示，移动学习关键技术首先需要考虑构建移动学习所需云、管、端系统支撑环境，在此支撑环境的基础之上，需要构建基于短视频移动学习资源的个性化智能学习环境。更进一步，移动关键技术涉及“云、管、端、资源、服务”这 5 个相关领域，下面分别介绍上述相关领域的国际研究现状。

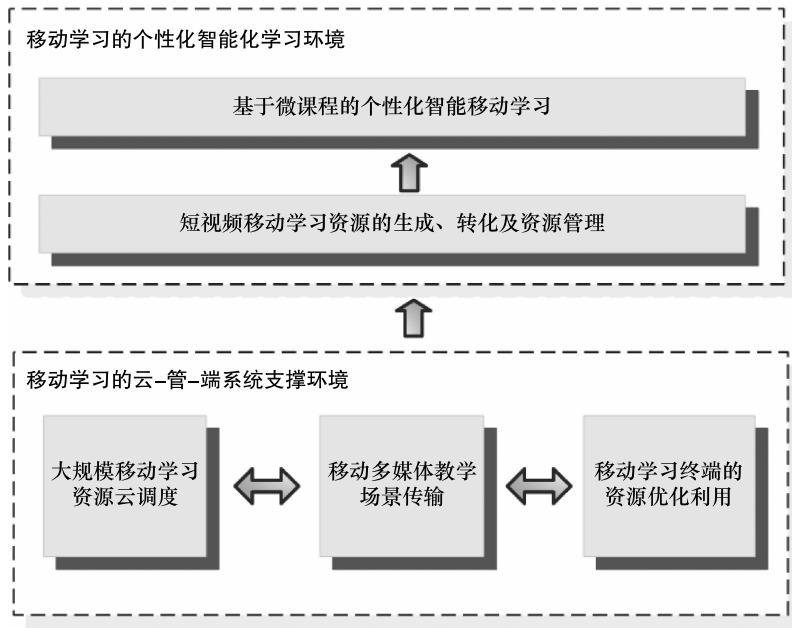


图 1 移动学习核心关键技术涉及的研究领域

1. 基于云计算虚拟化环境的大规模移动学习资源调度技术领域

基于云计算虚拟化环境的大规模移动学习资源调度技术的核心是面向规模化的多媒体移动学习服务，考虑移动学习的规模化、动态化特征，实现多媒体云计算平台资源的高效利用。国际上的相关研究主要涉及虚拟化云计算资源的调度及多媒体云计算服务。

(1) 云计算虚拟化环境下移动学习资源调度

云计算虚拟化环境下移动学习资源调度主要涉及云计算环境中虚拟机资源的分配问题，更多的被看做虚拟机的放置问题。它通常被描述为多维向量装箱问题^[4]。优化的目标主要有：资源利用率、网络开销、能耗、迁移次数等；主要考虑的约束条件有服务级别协议、应用相关性等；场景包括动态放置与一次性静态放置，主要通过启发式贪婪算法求得近似最优解。例如，pMapper^[5]系统采用一种扩展的首次适应降序算法（First-Fit Decreasing, FFD）算法，降低了虚拟机放置过程中由于迁移所导致的能源耗费。研究^[6]为了降低物理机资源占用，提出了一种基于最大最小蚁群算法（MAX-MIN Ant System, MMAS）的元启发式算法。研究^[8]为每个物理机设定期望CPU和内存使用率，通过实时监测以事件触发被动实现虚拟机的重新放置。此外，粒子群优化算法（Particle Swarm Optimization, PSO）、遗传算法（Genetic Algorithm, GA）、蚁群算法（Ant Colony Optimization, ACO）等进化算法也被用于解决单目标及多目标的虚拟机放置问题^{[5][6][7]}。研究^[10]同时考虑数据中心资源利用率、能耗与热耗散的多目标优化过程，并采用一种带有模糊多目标评价的GA算法求解。研究^[7]采用改进的ACO算法来实现最少资源利用率与能耗的虚拟机放置。

此外，由于移动学习的短暂性、间歇性，可能会造成云计算资源的浪费。因此，虚拟机资源配置的动态分配与调整，也是移动学习技术涉及的领域。在国际上，许多研究工作根据应用负载的变化，为虚拟机动态分配资源。Padala^[8]为虚拟机设定优先级，并使用单输入输出控制器依据实际CPU利用率和期望CPU利用率之间的差异进行分配，在发生资源竞争时优先保证高优先级应用的资源需求。研究^[9]提出使用自适应多变量控制器，协调多层次应用中各层的资源分配以满足应用服务等级协议。研究^[10]使用反馈控制方法，依据实际内存利用率和期望内存利用率的差异对内存资源进行动态分配。研究^[11]提出使用在线前向预测和反馈控制相结合的方式对CPU资源进行分配。研究^[12]结合预测的思想，预测应用在未来的响应时间，把资源分配问题抽象为非线性优化问题。研究^[13]使用模糊模型和模糊预测方法为每台虚拟机设计局部控制器，自动学习虚拟机运行时的行为。

(2) 多媒体移动学习云服务

从多媒体移动学习云服务这个更高的服务层次来看，多媒体移动学习云服务领域的研究主要分为两类：1) 面向多媒体移动学习云服务的资源管理。通过优化云端资源来提升多媒体服务，如云端多媒体资源共享、面向多媒体服务质量保障的云计算资源分配等。2) 基于云计算的移动学习媒体计算。利用云计算的并行计算处理能力进行移动学习多媒体计算，如面向移动学习资源的视频转码、视频检索、多媒体数据挖掘等。

一方面，对于面向多媒体云服务的资源管理研究，国际上的最新进展是以流媒体点播等多媒体云服务为对象，以多媒体云计算平台的可伸缩与按需服务特性为立足点，从资源共享与带宽分配^[14,15]、缓存策略^[16,17]等角度出发，研究多媒体云计算平台的系统资源优化利用问题；另一方面，基于云计算的媒体计算研究主要针对多媒体处理应用，以云计算的并行处理能力为立足点，研究如何高效的完成媒体转码^[18~20]、媒体检索^[21,22]

等媒体计算任务，通过负载均衡等手段，优化多媒体云计算平台的资源利用。例如，针对视频转码。

在上述国际研究中，研究成果虽然未直接应用于移动学习系统，却是解决移动学习高质量化、资源微小化、学习间歇化的有效途径。但是，国内外现存的多媒体云计算领域相关的研究工作存在“平台”与“应用”严重脱节的问题。从本报告的角度出发，深度融合虚拟化技术与多媒体移动学习云服务，是移动学习技术所面对的挑战与机遇。

2. 大规模高质量移动多媒体教学场景传输技术领域

将教师授课现场以直播的方式在移动用户终端呈现，或是录制成多媒体课件，以点播的方式随时进行回看，是移动学习系统中主要的学习手段，其技术难点在于如何在移动网络中进行大规模、高质量的实时多媒体数据传输，从质量与规模两个角度应对移动学习的需求。下面介绍国际上本技术领域的研究现状。

(1) 高质量移动多媒体教学场景传输

首先，需要面对丢包区分问题。面向有线网络环境的传输协议，认为产生丢包的原因因为网络拥塞，进而改变发送速率进行拥塞控制。在移动学习的无线、移动网络环境中，当数据在无线信道中传输时，除了链路拥塞会造成数据包丢失以外，还存在由于路径损耗、多径传播与衰落和噪声干扰等原因造成的链路差错丢包。因此，要想实现高质量的移动多媒体教学场景传输，以解决丢包区分问题入手是一个研究思路。

针对这一问题，在国际上，研究者提出了以数据传输参数特征为依据的区分丢包原因的方法，并以此为基础对传输协议进行优化。其中文献^[30]提出的面向实时多媒体应用的移动网络丢包区分算法，其核心工作是以双向传输时延（RTT）区分丢包产生的原因，对基于有线网络的流媒体数据传输拥塞控制算进行改进。

其次，需要面对差错控制与拥塞控制问题。一般来说，实时多媒体传输 QoS 优化研究可分为差错控制与拥塞控制两大类问题。

差错控制方法可分为前向纠错编码、请求重传和错误隐藏。前向纠错编码通过添加额外的冗余信息，使得在数据包出现丢失的情况下，接收端仍然能够重构出原始数据信息。在有线网络环境下，由于网络性能相对稳定，丢包率及所需的编码增益相对固定。在移动网络环境下，网络性能波动频繁且相对剧烈，使得固定编码增益的前向纠错编码效率明显降低。对此，在国际上，一些研究^[23,24]提出了自适应前向纠错编码，其核心思想是依据移动网络状态的变化而动态地调整编码增益，使得错误恢复能力与链路丢包情况相匹配，从而降低额外冗余数据产生的带宽消耗；请求重传由接收端反馈数据包丢失情况，请求发送端重新发送丢失的数据包。对于视频点播类应用，一般采用延时约束的重传机制^[25]，以避免重传数据到达接收端后由于落后于当前播放时间点而被丢弃造成的带宽浪费。对于实时性要求更高的视频直播类应用，一般认为重传技术并不适用；错误隐藏在丢包已经发生的情况下，视频接收端通过消隐的方式隐藏丢失的视频数据信息。错误隐藏有两种基本的方法，即空间插补和时间插补。空间插补^[26]以邻居空间的像素值来重构丢失的像素值，主要应用在帧内编码时丢失数据的重构。时间插补^[27]以前一帧的视频数据来重构丢失的视频数据，主要应用在帧间编码时丢失数据的重构。错误隐藏属

于被动性差错控制方法，不会增加额外的数据传输开销，但增加了移动终端的解码计算复杂度。

拥塞控制方法的核心思想为通过调整发送速率与链路吞吐率一致，避免出现由于发送速率大于链路吞吐率而造成链路拥塞。拥塞控制可在发送端或接收端实现。基于发送端的拥塞控制方法，通过探测当前链路的可用带宽，调整发送速率以避免出现拥塞。链路带宽的检测可分为探测方法和模型方法。速率调整可分为“加性增，乘性减”和“乘性增，乘性减”。

最后，需要研究移动学习教学场景传输网络模拟测试平台。针对移动网络的各种传输优化算法，最终需要移动网络模拟测试平台进行算法评估、对比与改进，为高质量移动多媒体教学场景传输解决方案的测试与验证奠定基础。

以 4G 移动网络为例，由于关注的领域不同，在国际上，LTE 模拟测试软件之间的功能存在差异。例如，文献^[28,29]主要实现了针对 LTE 物理层的模拟；文献^[30]模拟了 LTE 链路层的特征；文献^[31]模拟了 LTE 中的多输入多输出（MIMO）功能。基于 LTE 网络开展移动学习应用研究，如实时多媒体教学场景传输 QOS 优化，则需要 LTE 模拟测试软件能够模拟完整的 LTE 演进型分组核心网（Evolved Packet Core Network）功能与协议栈，如文献^[32~34]中的工作。

（2）大规模移动多媒体教学场景传输

支持异构终端与网络的大规模、多元化移动多媒体传输方法，也是大规模移动教学场景直播、交互必须要面对的挑战。其紧密相关的技术就是大规模流媒体传输技术。

大规模流媒体传输技术经历了由单播（Unicast）、IP 组播（IP multicast）到内容分发网络（Content Delivery Network，CDN）、P2P 流媒体传输、多媒体云计算（Multimedia Cloud Computing）的演变过程。其中，CDN 与 P2P 流媒体传输技术分别由服务器和用户端系统在 Internet 上构建逻辑独立的覆盖网络（overlay network），基于覆盖多播（overlay multicast）实现流媒体数据的分发。与 P2P 流媒体传输技术相比，CDN 技术具有可控、可靠的优点。因此，CDN 技术多应用于对服务质量有较高需求的增值服务领域。例如，基于 CDN 面向远程教育的多频道直播与共享式交互技术^[35,36]。与 CDN 技术相比，P2P 流媒体传输技术具有成本低、易部署、可扩展性强的显著优势，适于超大规模的多媒体应用。一些与移动学习密切相关的 Mobile P2P 研究成果^[37]，可为移动学习系统的设计借鉴。近年来，在国际上，基于虚拟化的云计算技术，因其具有良好的可伸缩特性与易部署、低成本优势，已成为在 Internet 上部署大规模多媒体应用最有效的方式之一，并可与前述 CDN 及 P2P 技术有机结合^[38]。

3. 移动学习终端的节能与资源优化利用技术领域

多媒体移动学习终端作为一种特殊的视音频传输的应用终端，具有持续时间长、能耗高的特点，移动学习终端的节能与资源优化利用至关重要。移动终端的节能主要包括硬件层、操作系统层和应用层三个方面。作为移动学习服务的提供者，应用层是系统设计与构造中可控的层面，因此，我们重点从应用层的角度来了解其省电设计方法。

对于移动学习终端来讲，如果终端的服务采用 B/S 架构，那么则可以对终端的浏览

器的能耗进行优化。例如，Bo Zhao 等人提出了一种能源感知的智能浏览器^[39]，当浏览器载入 Web 应用时，浏览器能够智能地将应用中的计算按照对能源需求的多少降序排列，这样使得浏览器运行完能源需求高的计算之后，降低 3G/4G 到较低的电能模态，执行余下的计算。这样有效地排除了浏览器执行非顺序计算序列所导致的电能模态无法在充足的时间内向低模态转换的可能。Bo Zhao 等人还提出了一种更为直接的方法^[40]：使用代理服务器将移动终端中网页浏览器的计算转移到代理服务器中，这样从根本上减少了移动终端的计算量，从而达到降低计算延迟、减少计算产生的能耗的问题。

对于移动学习中的多媒体传输（例如视、音频传输）这种持续性强、能耗更高的应用来讲，通过一些应用层的技术则可以增加终端在服务过程中转为空闲状态的可能。例如，Matti Slepkinen 等人在多媒体数据传输的过程中使用代理服务器进行速率整形^[41]，将持续发送的媒体流变为一段段突发数据簇，通过优化参数（如模态转换的时间阈值、突发数据簇的间隔时间等），使得电能模态可以在突发数据簇的间隔时间内从高模态向低模态转换。经测试，该方法在大多数移动终端的多媒体应用中能有效减少电能的消耗，可供多媒体移动学习技术借鉴。

对于移动应用，特别是多媒体数据传输应用，还有一种可行的方案是以提高能源效率为导向，对传输策略进行重新设计。文章^[42]介绍了一种节能传输算法，该算法根据信道环境，动态改变编码方式和传输策略，以降低能量的消耗。

此外，移动学习系统功能的多样性也需要在节电管理中考虑，如很多的移动学习终端在视频直播的过程中会有大量频繁的师生交互行为。在这种条件下，对于交互行为进行测量建模甚至预测，并根据预测结果动态地限制和调整手机功能，在不对用户体验造成很大影响的情况下降低能耗便成为一个很大的挑战。可以借鉴文章^[43]中对移动终端游戏应用的行为进行建模的方法，对移动学习应用进行类似的研究。

4. 短视频移动学习资源的生成、转化及资源管理技术领域

短视频是移动学习资源的重要组成部分，是移动学习过程中知识传播的主体，相对于文本、图像和音频等资源，短视频能够以最直观最接近自然授课的方式向学习者传播知识。根据 MIT 对 MOOC 教学视频的研究，短视频能够显著提高学生参与学习活动的程度^[44]。而且，短视频易于分享传播，可以与知识点、知识地图形成对应，便于导航和检索，学习者可以充分利用碎片化时间进行学习，教师也易于形成独特的教学风格。

为了充分利用短视频学习资源的诸多优点，移动学习技术需要在其生成、转换与管理方面进行技术研究与实现，以下是当前这一技术领域的国际研究现状。

（1）移动学习的短视频资源生成与转换技术

目前国际上对于移动学习的短视频学习资源生成的研究主要包括两个方面，分别是视频教学效果研究和采编技术研究。对于前者，edx 在此方面发表了研究报告，提出了一些安排教学视频的指导意见，包括画面节奏与图像编排、模拟课堂板书、语速语调、观看体验、操作过程等诸多方面。对于后者，目前视频的采编录制方式主要包括内录式和外录式^[45]，内录式是利用录屏软件自动录制教师讲课时的计算机屏幕（大部分为 PPT），录制教师声音，有时也会录制教师影像。外录式是利用专业摄像团队对教师上课场景进

行录制和后期制作。

除了移动学习短视频资源的生成外，短视频学习资源转化技术也十分重要，现有的大部分视频学习资源是过去录制的以现实课堂为单位的长约 40~120 分钟左右的长视频。把现有的视频转化为面向移动学习的短视频，是解决移动学习资源短缺的有效途径。短视频的转换方法可以归结为两种。一种是标记视频知识点与时间戳的对应关系，使用视频切割软件进行定位切割。另外一种是根据视频内容，使用视频场景分析算法进行智能的视频切割。

首先，对于有标识的知识点与时间戳的学习视频资源，可以使用视频切割软件根据知识点对视频进行切割，形成短视频学习资源。由于无知识点标识视频学习资源中大量为 PPT 的展示配上教师的声音，因此可以使用视频分析技术，对视频内容进行分析，找出知识点并进行切割，但其涉及图像识别与数据挖掘技术，具有较大的难度，依然是一个值得研究的开放性挑战。

综上，移动学习短视频资源生成与转换技术中，其主要难点或主要的关注点并不是具体实现方式本身，而是在于结合认知科学、教育学与人机交互等领域的经验，设计符合移动学习特点的教学资源。

(2) 短视频学习资源存储与管理技术

短视频在其存储形式上，通常表现为占用磁盘空间较少的小文件，对于如何存储和管理这种小文件，国际上有以下相关的研究。在移动学习短视频文件的存储管理上，利用其文件较小的特点，可以采用多种优化小文件存储的技术和策略来优化其存储结构。利用文件在系统内的逻辑关联性，可以将经常访问的文件集中在一起合并存储，这样在读取时就可以对相关文件进行预取，减少对磁盘的 I/O 操作，提高读取效率。另外，也有利用 Memcached 来提高小文件系统的读取吞吐率的方法，采用 Memcached 系统来为文件系统构建缓存系统，解决磁盘 I/O 性能对小文件读取过程的制约^[46]。

总的来说，主要是通过利用多级的缓存来加速小文件的访问速度，从而提高其读取性能。在具体的文件管理方面，则可以采用文件系统或者分布式文件系统进行存储与管理。

5. 基于微课程的个性化智能移动学习技术领域

移动学习的发展对传统的学习模式提出了新的挑战。目前大部分移动学习资源只是传统数字化学习资源的简单迁移，存在内容重复、质量低、设备兼容性差等问题，无法满足移动学习的需要。移动学习要求课程篇幅短小、能兼容不同便携设备，可以充分利用学习者的碎片时间、可进行随时随地的学习，即学习者可利用生活中零散的时间随时随地开展学习，这种需求让越来越多的研究者关注到了微型课程。

(1) 微课与微课平台

微课（Microlecture）指运用建构主义方法、以在线学习或移动学习为目的的实际教学内容。在国际上，它由美国新墨西哥州圣胡安学院的戴维·彭罗斯（David Penrose）于 2008 年首次提出。简短的微课视频让学生能更好地集中注意力把握课程要点^[47]。微课不仅可用于在线教学、混合式教学、远程教学等，也为学生提供了自主学习的资源，

让学生随时随地进行知识巩固学习。

微课平台用于微课资源、学习者及微课发布者的组织与管理，包括微课资源的存储、传输及可视化，学习者访问权限、学习记录及基本测评等管理功能。国外最具影响力的微课平台包括可汗学院（Khan Academy）及 TED-Ed，其微课视频发布于 YouTube 平台。

（2）基于微课的个性化移动学习

移动学习是当前碎片化和泛在学习时代获取信息的新型学习模式，具有普适化、情境化和个性化的特征，能实现校内、户外和家庭的无缝化持续学习，为混合学习、社会化学习、泛在学习等新的学习模式提供技术支撑，进而激发新的学习模式出现，给人们的学习方式带来革命性的变化。

与传统的数字化学习（e-Learning）相比，在国际上，移动学习的个性化研究呈现出如下新特点^[48]：

- 学习者所处的情境变化很快（如位置、周围人群、周围干扰等的变化）。
- 移动学习终端设备类型多样，如手机、PDA、平板电脑、电子书包、专用学习设备等，造成学习者数据来源复杂性高，存在文本、视频、语音等多种模式。而且设备本身的计算能力等会受到约束。
- 学习服务需要更高的时效性，要根据情境变化和用户的兴趣快速更新。
- 推翻传统课题的程式化教学，不但需要学生，而且需要教师具备新的设计理念。

如何自动感知移动学习者的情境、学习内容、社会信息等特征，为学习者推荐符合其所处情境和时效性强的学习服务，已成为移动学习中亟待解决的关键科学问题之一^[48]。

基于以上新特点，不同于传统的 e-learning，现有的移动学习更加强调“在合适的时间、合适的地点、以合适的方式把正确的内容传递给不同的学习者”。所有这些都直指个性化的两个关键研究内容：用户建模和个性化推荐算法。

1) 用户建模

移动学习中用户建模的实质是对同移动学习环境交互的用户进行跨学科分析研究。相关国际研究中，它涉及用户基本信息、兴趣（需求）发现、上下文（情境）、情感状态、社会信息以及相关标准研究。

在基本信息（Profile）表示方面，学习者模型在教育自适应超媒体和教育用户模型研究领域已被广泛研究。Brusilovsky 和 Millan、Specht、Nguyen 和 Do 提出的学习者的主要特点包括：个人基本信息、知识水平、兴趣、学习目标、学习和认知风格、情感、用户背景^[48]。

在情感状态建模和使用方面，情感信息的建模和使用在不同的研究领域都是热门的研究课题。麻省理工学院媒体实验室在此方面做了很多工作。例如，罗素^[49]的二维“环形情感模型”，该模型中人的情绪被看做是觉醒和价态的组合；OCC^[50]的模型也被广泛引用，这个模型指定 22 个情感类别。毫无疑问，情感同学习者的学习效能有紧密联系。

在上下文（即情境）信息方面，上下文被认为是移动学习用户建模中区别于传统学习的重要因素。目前基于上下文感知的推荐研究是国际研究的热点。现有的研究从上下

文类别可分为位置上下文、计算上下文、时间上下文、物理上下文、活动上下文（如拓展情境化元数据）、资源上下文等。

在兴趣发现方面，移动学习中的学习者兴趣感知和发现研究主要解决如何从不同情境、不同交互和不同设备等多个角度感知和发现学习者的兴趣。在个性化兴趣模型的构建方面，有基于关键词向量、基于概念、基于语义网等方法，但这些研究多侧重于知识间固有语义关系的分析，对学习者自身的学习能力、学习需求以及已有的学习经验等方面的因素没有加以考虑，即没有考虑移动学习者自身的影响因素。

在移动用户模型信息获取与发现方面，对移动学习者模型信息的获取主要有显式获取和隐式获取两种方式。显式获取就是引导用户主动提供信息，隐式获取是通过用户数据挖掘技术得到。目前大多数研究者采用隐式获取方式。学习者获取和发现领域研究一般采用机器学习、数据挖掘等技术，如聚类、决策树、贝叶斯分类等。

在对于移动学习者群体的兴趣感知和发现方面，国际研究者提出基于移动用户行为^[51]、结合个体用户需求和移动上下文信息的方法等^[52]。

在社会信息的描述方面，社会关系描述两个或两个以上的人之间的社会团体，连接或隶属关系。例如，社会关系可以包含有关朋友、中立者、敌人、邻居、同事和亲戚的信息。其他的研究人员把社区看做是一种重要的上下文尺寸。在 TEL 应用程序中，有时专家、教师和同龄学习者之间也有区别。

在学习者数据模型标准方面，学习者数据的部分表示的相关标准和规范是 IMS LIP、IMS Portfolio、IMS Enterprise、IEEE RCD、FOAF 和 HR-XML^[48]。对这些标准中数据元素的详细讨论超出了本报告的范围。Dolog and Nejdl^[53]讨论了这个领域不少有趣的工作。

2) 个性化推荐算法

面向移动学习的个性化推荐涉及具体的移动学习推荐及推荐算法评价两个方法。

传统的个性化推荐算法按照前面采集到的用户模型，结合某种过滤算法，完成对学习内容、学习策略、学习资源等方面的推送。对过滤算法最广泛被采用的分类是：①协同过滤，②用户信息统计过滤（demographic filtering），③基于内容的过滤，④混合过滤。协同过滤方法是一种常用方法，它利用一个用户同其他用户对资源（或者学习材料）集合上评分的共性特征向其推荐。用户信息统计过滤方法主要以某一共有的用户属性（例如性别、年龄、国家等）来判断是否有共同的偏好。基于内容的过滤利用学习者过去的学习内容进行推荐，可分为基于项目的和基于用户的。混合过滤通常是将用户信息统计过滤方法和协同过滤方法结合起来使用。

鉴于移动学习中上下文被认为是区别于传统推荐系统的首要因素，因此基于上下文的推荐成为个性化学习推荐中的研究热点。当前的移动学习的推荐研究，均利用上下文信息同各种算法形成不同的组合形式，大体上分为三类：基于上下文驱动的查询和搜索、上下文预过滤^[48]和上下文建模的推荐。

Santos 等人比较了不同的上下文推荐算法的组合的性能。当然，推荐系统特别是基于协同过滤算法和基于内容的推荐算法都伴随着冷启动、过拟合等问题^[48,54]。

个性化移动学习技术的核心就是基于移动上下文信息，实现个性化的学习资源推荐。

对评价移动学习个性化推荐的效果，目前主要是从学习效率和/或学习效益、准确度、有用性和可用性等方面评价，文献^[48]对此进行了详细阐述。

综上，基于移动学习的特点，移动学习中基于学习者上下文信息的感知和发现研究将朝向多终端、多源、多模式用户数据的融合分析发展。这些都依赖于传感器，目前还没有很好地研究上下文信息的传感器采集的粒度和尺度。

2.3 移动学习系统的部署应用

在移动通信技术和无线网络技术飞速发展的今天，随着智能手机、PDA、平板电脑等移动终端的普及，用于支撑起移动学习这种新型学习方式的移动学习平台、系统的研发和应用呈现出蓬勃发展的趋势。移动学习系统部署应用的发起者主要分为两类，一类是传统的 e-learning 系统供应商，他们想借助数字学习的经验，将传统 e-learning 系统改进成支撑移动学习的系统和平台，希望将移动学习推向市场，用于企业培训、网络学习；另一类则是教育科研机构，他们希望借助先进的移动技术和移动学习理论来改善教学环境，提高教学质量甚至解决教育资源分布不均的社会问题。

国外移动学习系统的部署应用呈现迅速发展的趋势，尤其是欧洲、北美等经济发达地区，由政府、科研机构以及企业发起的移动学习研究项目就超过 100 个^[55]，相应的移动学习系统的应用也全面展开，其领域分布于 k-12、高等教育、职业教育、远程教育、企业培训以及泛社会化的大众教育。比较有影响力 的项目有英国 Ultrallab 实验室的 M-learning、意大利 Giorgio Da Bormida 负责的 MOBILearn、爱尔兰远程教育专家 Desmond Keegan 负责的 From e-learning to m-learning 和 Mobile Learning The Next Generation Of Learning^[56]等。根据不同研究目的和应用领域，在这些项目的实施中都研制了不同的移动学习系统，比如在芬兰进行的 UniWap 项目中使用的移动学习系统通过 WAP 技术，开创了基于 SMS 短消息、MMS 彩信的移动学习模式；日本德岛大学开发的 BSUL 环境，通过上课使用无线 PDA 来及时地反馈信息，以达到教学互动的效果；芬兰 Tampere 大学开发出 XTask 移动学习系统用于协作学习；瑞典 Vaxjo 大学开发出 C-Notes 系统，能有效支持在无线环境下协作构建知识库；欧洲 M-learning 项目开发的 MediaBoard^[57]系统和英国诺丁汉大学开发的 MyArtSpace^[58]平台可以支持社交化的移动学习，学习者之间可以分享学习资源、学习体会。

通过以上项目中移动学习系统的应用，可以看出早期国外移动学习系统功能比较简单，使用的规模也不大，其应用的主要目的在于进行移动学习的理论、模式、效果的探索和验证；如今国外移动学习系统融入了更多的新技术，其功能完善，支持的终端设备也更多，服务的规模也越来越大，服务的内容更丰富，要求的质量也更高，应用的范围也更广。

从市场的角度看移动学习系统的应用，Ambient Insight 的最新报告显示，在全球范围内，移动学习产品和服务的规模将从 2012 年的 53 亿美金增长到 2017 年的 122 亿美金，在这 5 年内的复合增长率将达到 18.2%。在 2012 年，美国是最大的移动学习买方市场，

其次是亚洲地区^[59]。ASTD 的研究报告显示，2012 年，美国有 57.2% 的组织计划在未来三年内设计能够在移动设备上学习的内容，专注于移动学习近 15 年的美国移动学习专家 Chad Udell 预估，到 2015 年，几乎所有美国的企业都会使用移动学习系统或移动绩效支持服务^[60]。从这些咨询报告来看，美国的移动学习系统的应用已经进入比较成熟的阶段。

从产品的角度看移动学习系统的应用，在国外与移动学习系统的相关的产品非常丰富，从美国数字化学习产业网的调查可以看到，用于支持移动学习的内容创作和分发工具平台就超过 39 个^[61]。由于移动设备的屏幕尺寸、兼容性不同导致制作和分发移动学习内容相对困难，而 Adrenna Mobile 工具就可以解决这个问题，它可以支持正式的、非正式的、社交的和协作的学习，所有主流操作系统都可以使用这个工具。在国外，提供移动学习解决方案的公司也非常多，其中以 BlackBoard 较为有名，作为全球最大的网络学习方案服务商之一，该公司的产品几乎覆盖了所有对网络学习有需求的领域，比如高等教育、k-12、政府、企业以及军队，其移动学习产品 BlackBoard Mobile 也成功地在哈佛、杜克等知名大学应用。CellCast 平台是由 OnPoint 公司开发的一个移动学习产品，该公司早在 2002 年就开始了移动学习产品的开发，其核心移动学习平台已经发展到了第六代。

3 国内研究进展

本章与国际研究现状章相对应，将对移动学习的国内研究进展，从移动学习系统架构、移动学习关键技术、移动学习系统应用三个角度展开分析与讨论。

3.1 移动学习系统架构的技术演变

我国移动学习领域的研究起步较晚，但移动学习系统架构技术的演变与国际上的类似，也是从早期的 C/S 或 B/S 架构演变为云端结合的架构。

早期的移动学习项目主要是在教育部的策划下开展，集中于构建校园局域网和基于短消息的移动教育研究。例如，教育部“移动教育”项目，该项目利用中国移动的 GPRS 平台向广大师生提供短信服务，同时让师生享受更加优惠的移动电话业务，建立“移动教育”服务站体系，为参与“移动教育”项目的用户提供各种服务。

随着云计算技术的发展，以云计算环境为系统服务平台，以智能手机、平板电脑等移动终端为应用终端的“云+端”结合的架构已成为大规模多元化移动式学习体系架构的发展趋势^[62]，如国内深圳问鼎资讯和北京捷库动力所提供的移动学习解决方案都是基于公有云的 SAAS 云计算模式。与之相似，国内更多的研究机构和高校等也在研发移动学习系统并积极推动基于云计算的移动学习系统的应用，如清华大学、西安交通大学等。

3.2 移动学习关键技术的研究进展

针对移动学习关键技术的国内研究进展，与移动学习技术国际研究现状相一致，分别具体介绍在“云、管、端、资源、服务”这5个移动学习技术领域的研究现状与趋势。

1. 基于云计算虚拟化环境的大规模移动学习资源调度技术领域

国内基于云计算虚拟化环境的大规模移动学习资源调度技术，相关的研究工作与国际研究现状相类似，依然是虚拟化资源的调度工作，主要是一些跟踪性的研究成果^[63-71]，包括基于资源利用率、能耗、迁移次数等的虚拟机部署与调度优化，主要考虑约束条件服务级别协议、应用相关性、及经济性因素等。

但是，国内的虚拟资源调度方法目前只是停留在理论和仿真层面的调度研究，对实际环境，特别是在开源的系统环境下，如基于OpenStack云计算资源管理平台的调度实证研究还有待加强。和国外相比，大规模移动学习资源调度技术还没有成熟的云计算资源调度实现可供参考。

2. 大规模高质量移动多媒体教学场景传输技术领域

在高质量移动多媒体教学场景传输的国内研究中，针对丢包区分问题，文献^[72]以相对单向传输时延联合平均丢包率，采用Fuzzy模型区分丢包产生的原因。在区分丢包产生原因的基础上，传输算法优化研究主要针对TCP协议及TCP协议变种进行优化研究，并未检索到针对实时多媒体应用的丢包区分问题的相关研究成果。

针对错误恢复问题的研究，在前向纠错编码领域，文献^[73]总结了FEC的基本原与丢包恢复技术的研究成果，并在此基础上提出一种基于RS码的自适应FEC丢包恢复的架构方案。针对实时多媒体数据的自适应前向纠错编码问题，文献^[74]提出了增强前向纠错算法EFEC，由接收端根据反馈信息动态调整冗余信息与编码增益，提高视频服务的传输质量；对于请求重传问题，文献[75, 76]分别从丢包对视频还原质量影响程度的角度，选择需要重传的数据。与移动网络相关的重传技术研究，主要针对链路层及以下的协议进行优化；对于错误隐藏问题，研究方向同样可分为基于空域和时域两大类。文献[77, 78]针对当前主流的H.264编码，针对错误对视频解码还原时产生的影响，采用不同的错误隐藏策略进行优化。并未检索到以计算复杂度为主要约束的错误消隐算法研究。

针对移动网络中拥塞问题，当前国内研究主要集中在针对标准TCP协议拥塞控制机制的改进^[79]和从公平角度出发的TCP友好型拥塞控制算法^[80]。以上研究的共同前提为丢包区分问题。

此外，在网络模拟测试平台方面，未检索到有开放可用的仿真测试工具。

另一方面，大规模移动多媒体教学场景传输的国内研究现状与国际研究现状基本一致。对于P2P流媒体传输方面的研究，根据网络架构可分为网状、树状和混合架构^[81]，根据数据交换方式可分为推送式、拉取式和推拉结合式^[82]。特别地，由于国内的网络带宽条件限制及规模化的用户特征，一段时间内，基于P2P的大规模流媒体传输应用发展

迅速，与国际研究相比，从实际应用的角度来看处于领先地位。在多媒体云计算方面，国内的研究刚刚起步，文献^[83]从移动终端多媒体业务类型的角度，研究拓展云平台的业务支撑范围。

3. 移动学习终端的节能与资源优化利用技术领域

在国内，一方面，移动学习终端的节能与资源优化利用技术的理论研究依旧是以移动终端的节电为主，但相关研究成果较少。Xin Li、Mian Dong 等人则通过用户的历史数据和当前的网络状况对于数据缓存进行动态管理^[84]，使得应用能够审慎地进行用户数据传输（包括多媒体的传输），减少了传输数据的冗余性。另一方面，国内研究者，特别是国内企业界，在定制的移动学习终端开发上，体现了我国在电子工业方面的优势，面向移动学习应用特点而定制开发的移动学习终端实现了硬件与移动学习应用软件的定制匹配，从而降低了移动学习终端的成本，从另一个角度实现了移动学习终端的资源优化利用。随着 Android 平板电脑成本的降低，多家高等学校的网络教育学院均推出了定制的移动学习终端。相关移动学习终端不仅可以提供传统移动学习所需的移动学习社区、直播、点播功能，还可以记录学习者的学习轨迹，为网络教育学院提供教学评价，以及提供个性化服务。

4. 短视频移动学习资源的生成、转化及资源管理技术领域

国内针对短视频学习资源的生成技术方面的研究主要集中在对视频内容的设计和过程编排上，除此以外主要是利用已有的视频剪辑与后期处理技术对拍摄完成的视频进行处理。而在短视频学习资源的转化与资源管理方面，国内均有相当数量的研究。以下是对短视频学习资源的转化与资源管理方面研究进展说明。

在短视频学习资源的转化方面，随着云计算技术的广泛使用，视频转码也逐渐开始采用“云端转码”进行实现。传统转码方案依赖单一服务器，存在效率低和单点依赖问题，可以设计一种分布式流媒体计算框架实现分布式转码系统，基于 Hadoop 分布式文件系统和 Map Reduce 并行计算框架，以及 memcoder 技术，并利用负载均衡，可以有效降低了转码的处理时间，根据测试，采用该系统转码流媒体，处理时间可降低 70%^[85]。

在短视频学习资源的存储与管理技术方面，国内研究者在访问策略方面，将逻辑上连续的数据尽可能存储在物理磁盘的连续空间，使用 Cache 充当元数据服务器的角色并通过简化的文件信息节点提高 Cache 利用率，提高了小文件访问性能^[86]。或者利用数据库系统保存文件的元数据提高检索的速度，利用分布式文件系统保存文件内容，结合两者的优势来优化访问过程。在使用 Hadoop 分布式文件系统存储文件时，国内相关研究者，采用序列文件技术将小文件以队列的形式合并为大文件，实现了节省名称节点所占内存空间的目的，采用数据标准化方法和三精度层次分析法确定队列长度的最优值，使得小文件的合并能在合并时间、文件操作时间和节省内存空间之间达到一种平衡^[87]。或者利用数据库系统保存文件的元数据提高检索的速度，利用分布式文件系统保存文件内容，结合两者的优势来优化访问过程。或者利用 Redis 内存数据库来缓存常用数据，在保证不影响文件删除速度和随机获取文件速度的基础上，能够较大程度地提高内存的利

用率，加快文件的检索速度，保证了文件的快速、可靠操作^[88]。也可以采用基于内存的分布式文件系统存储段视频文件，基于内存的分布式文件系统能够利用内存的性能优势，提供高的数据吞吐率和低访问延迟。对于小文件存储系统，国内研究者发现其性能与基于磁盘的分布式文件系统相比具有明显的优势^[89]。在存储介质方面，可以利用 SSD (Solid-State Drive，固态硬盘) 进行数据读取的加速，利用 SSD 在随机读写带宽与 IOPS 方面的优势，能够明显提高元数据读取过程中的 I/O 请求吞吐率，在整体上提高服务响应速度^[90]。

5. 基于微课程的个性化智能移动学习技术领域

微课通常由简短的视频及配套资源组成，教师可在课堂上利用微课作为授课的素材，而学生可通过微课进行预、复习等，实现自主学习。

微课符合了网络时代学习碎片化的需要。微课内容的呈现形式多样，如卡通动画、电子黑板、真人演讲等，课程面向不同年龄、专业的人群，其内容短小精悍，时长一般在 10 分钟左右，并配有对应的字幕便于学习。微课与传统的课程有着较大的区别。从授课时间来看，微课一般是 5~10 分钟，而常规课程一般是 40 分钟甚至更长；微课以知识点为对象，而常规课程一般包含多个知识点。

国内的微课平台包括中国微课网 (cnweike.cn)、微课网 (vko.cn) 及网易开放课程平台 (<http://open.163.com/>) 等。其中，中国微课网目前涵盖了来自全国 31 个省市的中小学教师上传的参赛微课视频，涉及语文、数学、英语等，学科授课时长均在 10 分钟以内。

另一方面，在个性化移动学习方面，国内很多研究机构已经意识到上下文感知推荐系统^[91~93]的重要性，香港科技大学、浙江大学、西北工业大学、北京邮电大学均开展了相关研究。王立才等人^[92]详细地综述国外上下文感知推荐系统的框架、关键技术、主要模型、效用评价以及应用实践。

国内在学习者建模方面，用户基本信息，例如性别、年龄、职业、文化程度和收入水平等维度，被经常采用。在位置上下文方面，香港科技大学、浙江大学都做了充分的研究；在情感方面，田锋等人基于问卷调研，制定了网络学习环境下粗细粒度相结合的学习者的情感类别，并从中文交互文本中发现学习者的情感^[94]；在兴趣发现方面，国内学者也采用关键词向量、基于概念、基于语义网等手段^[95]。

在推荐算法方面，国内也广泛结合用户信息统计过滤方法，香港科技大学、浙江大学、西安交通大学等采用协同过滤算法及其改进算法。

总体上而言，在基于微课程的个性化智能移动学习技术领域，国内的研究基本以跟踪国外研究并改进为主。

3.3 移动学习系统的部署应用

我国的移动学习研究起步较晚，但面向不同层次、领域的移动学习系统的研发与应用的脚步从未停止过。近年来，国内部署了大量面向不同人群的移动学习系统。和国外

一样，早期由于移动通信技术水平低，移动终端设备的性能低下，用于移动学习研究和实践的系统功能单一、系统简单、稳定性差；但随着 3G、WIFI 等无线通信技术进步，以及云计算技术的发展，我国的移动学习系统的功能也逐步在完善，应用范围也在逐步扩展。

从企业与厂商的角度来看，根据移动学习咨询网的调查，自 2011 年以来，国内关于移动学习的 App 迅速增多，在教育培训、企业培训等各个领域与教育相关的 App 大约有 10 万款左右，这些终端应用程序大多都是传统 Web 学习系统的移动版。随着移动互联时代的到来，许多教育和企业培训机构为了方便自己的学生能随时随地的学习，纷纷将自己的在线学习系统移动化，比如新东方在线、沪江网校、中国电信网上大学、中传在线移动学习平台。从这些终端产品的功能来看，课程的学习是以下载离线学习和点播在线学习为主，学习过程的互动方式很单一，只是将传统的在线学习系统的前端由 PC 转移到了移动终端。

2012 年、2013 年先后在北京召开了移动学习开发者专业会议，从会议的内容可以看出，传统 e-learning 厂商都对移动学习表现出了极大的兴趣，都开发了适用于不同领域的移动学习系统和平台，比如深圳问鼎资讯提供的移动学习平台、北京捷库动力的移动学习平台、北京创想空间的全时系列产品。这些企业的产品主要应用于企业内部培训、教育培训机构，所提供的移动学习解决方案都是基于公有云的 SaaS 云计算模式，用户不必购买任何软件和硬件设备，就可以开展在线培训和移动教学，他们的运营模式也从卖软件、系统转变到卖服务，提供整体解决方案。

与企业厂商的解决方案不同，由高校与研究机构研发的移动学习系统在体系结构与功能上更加完整，并在高校中得到了广泛的应用。清华大学、西安交通大学、上海交通大学、浙江大学等都有实际的移动学习系统投入实际应用。

清华大学开发的学堂在线 MOOC 系统以 OpenEdx 为基础架构，目前已经能够提供基于 Android 的 APP 应用。浙江大学也开通了“浙江大学开放/移动”学习系统，该系统使用开源的 LMS 软件 sakai，并在该平台上进行了二次开发，改善了交互性，增强了易用性，其主要目的是辅助课堂教学。上海交通大学继续教育学院开发了基于 SJTU-OMR (open mobile real) 教学环境的 PPClass 在线课程直播系统，该系统中使用 vastcast (www.vastcast.us) 移动流媒体互动平台实现对移动学习的支持。

报告撰写团队早期研发了 SkyClass 天地网远程教育系统，这是我国第一套具有自主知识产权的天地网远程教育系统。目前，随着移动学习产业的迅猛发展，SkyClass 研究团队基于移动学习的“4A”特征（人人皆可，自主学习（Anyone）；时间碎片，随时可学（Anytime）；天地结合，无处不网（Anywhere）；各类终端，异构接入（Anydevice））研制出 SkyClass 移动学习系统。

SkyClass 移动学习系统针对移动学习的“4A”特点，建立了“云”、“端”结合的新型移动学习模式，提出后端为云计算支撑平台，前端支持各类移动终端、互联网终端，支持多模式接入服务的“云 + 端”移动课堂模式与平台体系架构。其整体架构如图 2 所示。

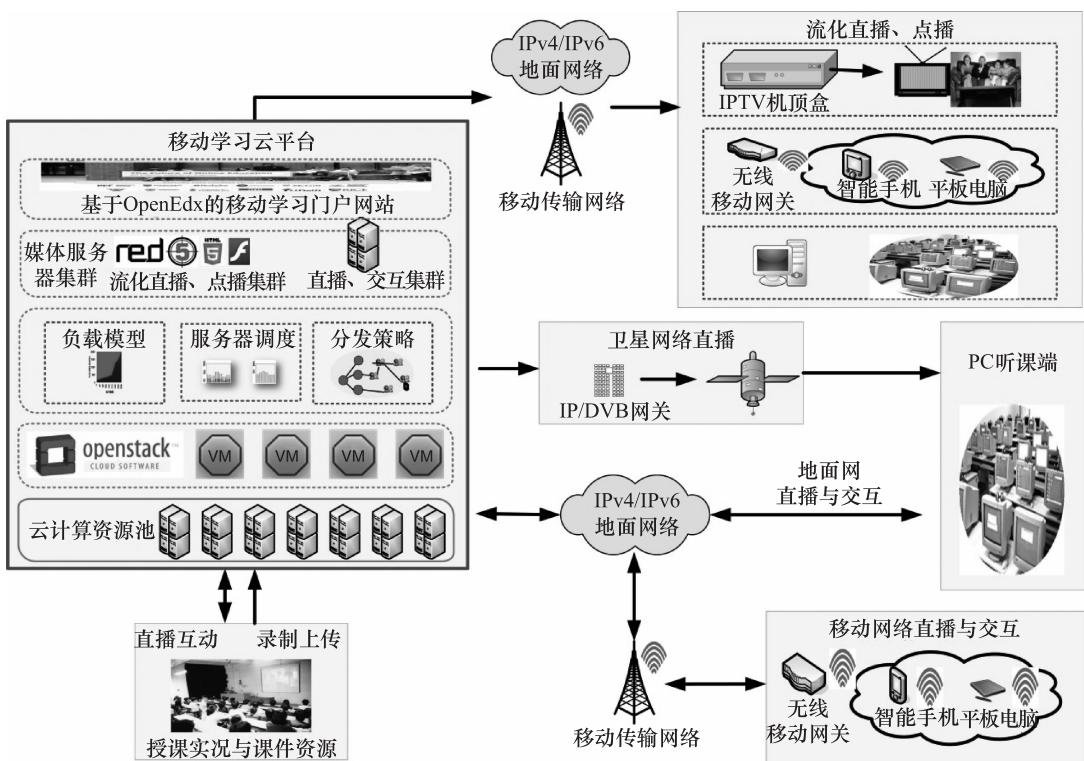


图 2 SkyClass 移动学习系统体系框架

SkyClass 移动学习系统提出从“云”、“管”、“端”三个方面优化移动学习模式的新思路，实现了以下关键技术：大规模移动多媒体直播点播云调度；移动网络多媒体传输 QoS 控制与优化；异构移动终端的多模式接入。

(1) 大规模移动多媒体直播点播云调度

云计算拥有良好的可伸缩性，如何动态调度云平台中虚拟机资源，使云平台资源利用率最优并能支持更多的移动请求是“云+端”移动学习模式面临的重大挑战。SkyClass 移动学习系统运用云计算开源解决方案 OpenStack 搭建私有云平台，动态监控、调度在其上部署的 Red5 流媒体服务器集群，以及直播交互集群。

(2) 移动网络多媒体传输 QoS 控制与优化

移动网络信道带宽具有不确定性以及数据传输的低可靠性，如何提高移动网络环境下实时视音频传输质量是移动学习技术亟待解决的重大难题之一。SkyClass 移动学习系统采用 NS3 以及 CMW500 进行移动网络传输特性挖掘，基于移动传输的特性和规律，从纠错编码和延时约束的选择性重传两方面优化 QoS 差错控制。

(3) 异构移动终端的多模式接入

基于移动终端的操作系统、分辨率、屏幕大小等环境各不相同，如何支持异构移动终端的无缝接入是大规模多元化移动学习技术面临的另一个挑战。SkyClass 移动学习系统采用 Red5 开源流媒体服务器，支持 RTMP、HLS 多种流媒体协议；基于分辨率、网络

状态、终端类型等因素为用户提供多种码率版本的视频流；客户端采用 Flash、HTML5 等播放方式实现对多操作系统、多浏览器的支持。

通过以上移动学习系统应用的实例可以看出，国内高校移动学习系统和平台的建设主要是围绕如何更好地服务本校的师生展开的，各校的网络学院、继续教育学院是移动学习的研究以及其系统研制的主要推动者，目的在于适应下一代远程教育的移动学习方式。

4 国内外研究进展比较

通过对移动学习的国外研究现状及国内研究进展的分析，可以发现在移动学习的体系架构、核心关键技术、系统与应用模式等不同领域，国内外研究进展各有不同，侧重也略有区别。整体上，国外移动学习的相关研究内容更丰富一些，但国内移动学习的应用更具多样性，国内外研究进展对比如表 1 所示。

表 1 国内外研究进展对比

类别	对比内容	国内研究进展	国外研究现状	备注
体系结构	移动学习体系结构的演变方式	由 B/S、C/S 至云端结合	由 B/S、C/S 至云端结合	国内跟踪国际演变
	云计算基础设施构成	主要依赖于私有云服务	比较完善的公有云解决方案，如亚马逊 EC2、S3，微软的 Azure，Google 的 AppEngine，以及 Akamai 的 CDN 服务等	国内缺少大家认同的成熟的云计算解决方案
核心技术	基于云计算虚拟化环境的大规模移动学习资源调度	有大量云计算虚拟化资源调度的研究成果可供借鉴	有部分云计算虚拟化资源调度的研究成果可供借鉴	均缺少多媒体云计算环境下媒体服务器集群资源协同调度研究
	大规模高质量移动多媒体教学场景传输	在高质量移动多媒体传输质量控制等领域处于领先地位	在 CDN、P2P 规模化服务等领域具有一定的应用优势	均缺少与移动教学场景紧密相关的移动多媒体传输方法
	移动学习终端的节能与资源优化利用	以通用的移动终端的节能优化方法研究为主	在推出定制的移动学习终端设备方面有优势	算法与硬件结合，是提升移动学习终端资源优化利用的有效途径
	短视频移动学习资源的生成、转化及资源管理	注重基于开源成果构建的生产环境	原型系统和实验环境下的理论研究为主	技术手段基本一致，核心的区别是移动学习资源内容的差别
	基于微课程的个性化智能移动学习		有部分基于微课程的知识地图导航学习成果	均缺少多终端、多源、多模式用户数据的融合分析

(续)

类别	对比内容	国内研究进展	国外研究现状	备注
系统与应用模式	移动学习系统的功能演进	从 SMS、MMS、WAP 到功能完备的移动学习系统	从 SMS、MMS、WAP 到功能完备的移动学习系统	国内外的移动学习系统都出现了与 MOOC 系统融合的趋势
	移动学习系统的运营推广	形成了完整的移动学习产业生态链	以研究性的系统应用为主，在商业领域的应用规模较小	国内移动学习系统部署应用规模有较快的增长
	移动学习系统的应用领域	涵盖了社会的各个领域	远程教育和考试培训	国的移动学习系统应用领域有较大的扩展

4.1 移动学习系统架构的研究进展比较

国内移动学习系统架构技术的演变路线基本上是沿着国际上移动学习系统架构技术的演变而演变的，都是从早期的 C/S 或 B/S 架构演变为云端结合的架构。然而，国内外移动学习系统架构的研究进展也有明显的区别，国外移动学习系统基本依赖于公有云服务。这是由于国外云计算发展较为迅速，有比较完善的公有云解决方案，如亚马逊 EC2、S3，微软的 Azure，Google 的 AppEngine，及 Akamai 的 CDN 服务等，国外的移动学习系统使用公有云提供的各种基础设施资源（如存储、CDN 等）来提供移动学习的服务。

国内移动学习系统往往依赖于私有云服务。一方面，由于国内云计算发展较慢，还没有大家认同的成熟的云计算解决方案；另一方面，由于国内各个机构的学习资源并不是共享式的，有些学习资源不便于公开。国内的移动学习系统往往需要建立“云”、“端”结合的移动学习系统，如包括“云”、“管”、“端”三个层次完整的解决方案，包括提供私有云端的资源管理和调度方法，移动网络传输优化方法和移动终端智能学习应用。

4.2 移动学习关键技术的研究进展比较

如表 1 所示，在移动学习关键技术领域，国内、外的研究方向与研究思路基本一致，但国内以跟踪国际研究进展为主。但在涉及大规模多媒体应用、移动学习终端设备等技术领域，国内由于有相关的规模应用需求与硬件设备环境，也有自身的一些优势。总体来说，国内外研究进展中，将移动学习特征深度融入相关技术领域的研究成果相对较少。

(1) 在基于云计算虚拟化环境的大规模移动学习资源调度技术领域

在国内外研究中，均有大量云计算虚拟化资源调度的研究成果可供借鉴。在国际上，更有如 OpenStack 这样的开源云计算平台管理软件可供实践。国内尽管研究成果相对较少，但也有部分云计算虚拟化资源调度的研究成果可供借鉴。基于云计算虚拟化环境的

大规模移动学习资源调度技术通常被描述为多维向量装箱问题。优化的目标主要有：资源利用率、网络开销、能耗、迁移次数等；主要考虑的约束条件有：服务级别协议、应用相关性等；场景包括动态放置与一次性静态放置，主要通过启发式贪婪算法求得近似最优解。然而，国内外的研究现状中，均缺少多媒体云计算环境下媒体服务器集群资源协同调度研究。

(2) 在大规模高质量移动多媒体教学场景传输技术领域

首先，通过对比发现，国内外在解决移动网络下高质量和大规模的视频传输优化问题的研究中，出发点与解决思路基本一致。例如，丢包区分研究是为了解决传统传输算法无法区分移动网络传输过程中丢包原因而造成效率降低的问题，解决思路一般是根据数据包到达数据接收端的规律，以传输延迟、延迟抖动和连续丢包长度等为量化参数，构建丢包特征模型，区分数据包丢失的原因。错误隐藏研究是为了解决在数据包丢失无法避免时，减少视频还原时质量降低对用户的影响，解决思路一般是在时域或空域对丢失的数据进行重构。在服务规模方面，已有研究均基于传统的 CDN 及 P2P 方式，或利用虚拟化的云计算技术的可伸缩特性与易部署、低成本优势，实现规模化服务。

其次，在移动网络仿真测试环境方面，国内研究滞后于国外研究，一般是对国外已有模拟器、工具进行改进或直接引用。在国内研究成果中，并未检索到能够支持端到端的多媒体数据传输优化的移动网络模拟工具。

最后，移动学习中的移动多媒体传输技术产生新的问题和挑战，主要原因是：①由于传输链路从相对稳定的有线/Wifi 环境转换为易受环境影响而产生波动的移动网络环境；②终端设备由 PC 机、笔记本等传统设备转换为计算、显示和续航等硬件配置与能力差异显著的异构移动终端设备两方面产生的。但是国内外缺少专门针对以上问题的研究，如移动网络在不同衰落场景下的传输特征的分析，针对移动终端计算能力的算法复杂度优化等。

(3) 在移动学习终端的节能与资源优化利用技术领域

国外研究现状是以通用的移动终端的节能优化方法研究为主，移动终端的节能主要包括硬件层、操作系统层和应用层三个方面。作为移动学习服务的提供者，应用层是系统设计与构造中可控的层面，国外研究现状中有较多的基于应用层的角度的移动终端省电设计方法。特别是，一些移动视频的节电策略可供多媒体移动学习终端的节能与资源优化利用所借鉴。国内仅有少量类似的跟踪研究。但国内研究进展中，在推出定制的移动学习终端设备方面有优势。有不少移动学习的服务与运营机构，包括一些国内高校的远程教育机构，均可提供定制的移动学习终端，体现了我国在电子工业方面的优势，从另一个角度实现了移动学习终端的资源优化利用。而节能算法与硬件的结合，也是提升移动学习终端资源优化利用的有效途径与发展方向。

(4) 在短视频移动学习资源的生成、转化及资源管理技术领域

在短视频移动学习资源的生成方面，国内外的研究并无显著差距，主要是在视频内容的编排与教学内容方面各有侧重。国外的短视频在教学内容方面，主要集中在技术思想演讲、基础概念教育以及高等教育方面。其中，技术思想演讲关注技术发展方向的引

导、创新思维的应用、热点问题的探讨以及各种工程问题；基础概念教育关注高中以下的数学、物理、化学、生物、历史等科目的一些困难知识点；高等教育方面则提供了计算机科学、电气工程、生物学、人文科学、哲学等多种学科的基础课和少量研究型课程。而国内目前短视频的内容主要关注职业技术教育、专业培训等方面，提供了包括计算机技术、财务、会计、工商管理等实用型的技能培训。随着 MOOC 在国内的发展和各大高校的加入，短视频中也逐渐增加了高等教育方面的内容，不断有相关课程资源生成。在技术手段方面，国内外并无明显差别，均采用内录式和外录式技术，并通过后期的视频剪辑形成最终的视频资源。

在短视频学习资源的转化方面，国内外均采用了基于开源技术的 Memcoder、FFmpeg 等，在云平台上进行视频资源的切割、转化与压缩等操作。Hadoop 的 HDFS 分布式存储系统和 MapReduce 并行计算框架是主要的技术手段。通常的转化目标包括多码率、多格式、多分辨率，将一个高清视频转化为适应多种屏幕分辨率以及多种播放器的形式，同时兼容宽带、无线网络以及移动网络，并通过切分视频提供面向多用户并发的点播支持。

在短视频学习资源的管理方面，国内外的研究方法基本一致，均是将短视频资源作为小文件的一种特例进行研究，从存储系统的层次划分中，为其设置合适的层次并进行存储优化与配置管理。在具体的实施方案研究上，国外有部分互联网公司提出了一些开源软件，在生产环境中实现了对短视频小文件的访问优化，国内在此方面主要以原型系统和实验环境下的理论研究为主，如何利用开源系统的生态环境，加快频移动学习资源的生成、转化及资源管理的实践，是国内研究者下一步应该重视的研究内容与领域。

(5) 在基于微课程的个性化智能移动学习技术领域

尽管我国研究者及教育家对微课程的研究及实践已取得一定成果，但与国外相比仍处于探索阶段。

一方面，在微课程的建设与设计方面，国内微课程的理论探讨、资源建设与教学应用仍需完善，国内外的研究进展比较如下：

首先，国内微课的表现形式较为单一。国内微课视频仍以课堂实录片段为主，其内容的连贯性不强，视频录制效果差，导致微课在教学上的应用效果欠佳。而国外微课内容的呈现形式相对丰富，除真人讲解演示外，还有卡通动画、电子黑板等形式，使得微课变得生动有趣，能引起学习者的兴趣。

其次，国内微课平台不够完善。国内微课平台目前只包含课程练习、交流、教案等基本功能，缺乏课程自定义、课程知识向导、数据统计、即时笔记等高级功能。在学习效果评价上更多采用客观习题进行量化评价，缺少对学生学习态度、教师教学效果等方面的评价。

最后，国内微课的应用不足。国外已将微课融入日常教学中，供学生进行自主预习、复习，并取得一定成效。而国内微课在该方面的应用研究较少，且对微课应用于教学的过程与效果分析、评价等方面的研究不多。缺少微课应用和学习体验的调查。

另一方面，在基于微课程的个性化智能服务方面，从用户建模和个性化推荐两个角度，对国内外研究进行比较如下：

在移动学习用户建模领域，首先，基于移动学习的特点，移动学习中基于学习者上下文信息的感知和发现研究将朝向多终端、多源、多模式用户数据的融合分析发展。这些都依赖于传感器，国外在传感器方面的研究处于领先地位，而我国相关技术的研究还处于起步阶段，这直接制约上下文信息的传感器采集的粒度和尺度研究。其次，国外虽然没有刻画移动学习用户上下文的统一规范和标准，但是已经有一些相关标准出现^[48]。而国内相关研究机构在参与此类标准的制定方面还存在一定困难。

在移动学习个性化推荐领域，首先，协同过滤算法及相关混合算法，依然在国内外研究中占据主要地位。但是最近几年，国际上基于知识的过滤研究逐渐兴起，因为它克服了协同过滤算法的冷启动、数据稀疏和过拟合等问题。国内也已经开始跟踪研究。其次，人类学习方式越来越明显地呈现出网络化、移动化等特点，国内学者已经指出：如何从人类自身的认知心理和行为特点着手开展对用户学习行为、兴趣感知的研究也是应该注意思考的问题^[96]。另外，周涛等在有关人类行为时间和空间统计规律的挖掘和建模研究综述中也提出：在人类行为时间特性的经典研究中，将泊松分布应用于人类活动量化模型中是存在不足的^[97]。这些研究和分析都为移动学习用户兴趣感知研究提供了新颖的理论依据和研究视角。

4.3 移动学习系统的部署应用比较

本节将对国内外移动学习系统的部署应用加以比较，通过对比可以看出，国内外移动学习系统的部署应用有相似的地方，但也存在着很大的不同。

从移动学习系统在实际研究项目中部署的阶段来看，国内外移动学习系统的应用都遵循着一个客观规律，首先是对移动学习的可行性及其理论的有效性的验证和探索的阶段；然后是将移动学习和各种学习模式结合的阶段，比如基于问题的学习、协作学习、非正式学习模式等；最后是将移动学习系统进行大规模推广的阶段。而从所部署的移动学习系统的功能和性能来看，国内外的移动学习系统也有着相似的发展过程。在早期由于通信技术和移动设备性能等因素的限制，移动学习系统功能都很弱，比如有基于 SMS 和 MMS 进行文本、图片推送的系统，基于 WAP 资源站点的在线学习系统等；而随着 3G、4G 以及 WIFI 技术的普及和推广，智能的移动设备将互联网装进了口袋，伴随着云计算、大数据的应用，移动学习系统在功能、性能上都得到了飞速的发展。

从移动学习系统的部署应用领域来看，国内外有很大差别。在国外，尤其是欧美等发达地区，移动学习系统所部署的领域涵盖了社会的各个领域，比如中小学教育、高等教育、社会教育、远程教育、职业培训等，这些投入应用的系统所涉及的研究内容也很丰富，如个性化教案设计、及时课堂反馈、虚拟社区学习、协作学习、游戏化学习、绩效支持等；在我国由于地区经济发展差距大，东西部教育资源不均衡，以及以应试教育为主的教育模式，移动学习系统的部署主要应用于远程教育和考试培训。

从移动学习系统的部署应用规模来看，国内与国外也有很大的差距。由于国外进行移动学习相关研究较早，传统的数字化学习系统提供商早就开始在移动互联网谋篇布局，

以美国为例，其已经形成了从移动学习终端应用、移动学习内容编制、分发工具到移动学习平台、内容建设的完整的产业生态链，占全球移动学习市场的份额最大，可以提供完整的移动学习解决方案的公司也更多；而我国的移动学习研究主要跟随国外，在理论研究方面有很多成果，但在移动学习系统的部署实施方面还欠缺很多，除了少数高校已进行的一些研究性的系统应用外，在商业领域的应用规模还无法与国外比较。随着国内移动学习基础设施（移动网络、云计算平台）的建设、移动互联网浪潮的来袭，我国的移动学习系统的部署应用规模将有较快的增长。

5 发展趋势与展望

通过国内外研究现状与进展的比较分析，移动学习技术无论从体系结构、关键技术，还是应用模式方面来看，都在快速发展与演变。掌握移动学习技术未来的发展方向与趋势至关重要。本节将首先简要归纳移动学习技术在体系架构与关键技术领域的发展趋势。随后，重点分析与展望 MOOC 与大数据技术对移动学习技术的推动与促进，介绍 MOOC 与大数据技术与移动学习技术的融合发展趋势。

5.1 移动学习的体系架构、关键技术及应用部署趋势

从移动学习系统架构的角度来讲，在移动网络技术与云计算技术的推动下，移动学习系统架构技术也在不断演变和进步，从早期的 C/S 或 B/S 架构演变为云端结合的架构，基于“云端结合、强云弱端”策略构建的移动学习体系架构是未来移动学习系统基础架构的必然趋势。而虚拟化云计算平台构建与移动学习应用情境感知相融合，综合云计算技术与多媒体移动学习服务的多媒体云计算研究，是移动学习体系架构研究未来的研究重点。

从移动学习关键技术的角度来讲，移动学习技术本质上是以移动学习应用为目标，由移动计算、云计算、网络多媒体、知识获取与个性化服务等领域相关技术支持而实现的一种复合技术。因此，对于移动学习关键技术的发展趋势，一方面，是以相关技术领域的发展为基础与依托，与相关领域技术的发展趋势相一致；另一方面，移动学习关键技术也有其特殊性与针对性，移动学习关键技术的发展趋势必然是移动学习自身特性与普适移动计算、云计算、网络多媒体、知识获取与个性化服务关键技术的紧密融合。例如，高分辨率、低帧率移动学习屏幕教案视频的传输方法，融合虚拟化技术与多媒体移动学习云服务的大规模移动学习资源调度方法，微课程短视频资源的生成、转换与存储管理方法，基于微课程导航学习与移动学习用户兴趣感知的个性化智能移动学习方法，以及移动学习终端的节能方法等，均是具有鲜明移动学习特征的典型移动学习技术，是移动学习技术未来的发展方向。下面具体就移动学习关键技术涉及的几个领域的发展趋势进行分析与展望。

(1) 在基于云计算虚拟化环境的大规模移动学习资源调度技术领域

在云计算环境下，移动学习的各种资源依托于云环境中的虚拟机、网络等资源对外服务，云计算虚拟化环境资源的动态优化调度是保证移动学习规模与质量的关键。

如前所述，一方面，云计算环境下的虚拟机调度技术国内外相关研究没有针对多媒体云服务的特征；另一方面，面向多媒体云服务的资源分配与管理，忽视了云计算平台的基础架构，从而导致两个方面的研究均未能全面掌握整个多媒体云计算平台的资源占用特性与规律，无法从全局优化平台资源利用。解决以上问题的核心与难点是深度融合虚拟化技术与多媒体云服务的研究工作。

因此，基于云计算虚拟化环境的大规模移动学习资源调度需要虚拟化云计算资源调度与多媒体移动学习应用情境感知相融合。一方面，需要考虑移动学习服务的高质量化、资源微小化、学习间歇化等服务特征；另一方面，需要考虑云计算环境的网络、计算、存储资源特征。而综合虚拟化技术与多媒体服务的多媒体云计算研究工作，是多媒体云计算环境下大规模移动学习资源调度技术的主要发展趋势。

(2) 在大规模高质量移动多媒体教学场景传输技术领域

在移动学习传输质量方面，以智能移动终端设备为载体，以移动网络为传输管道，其优化目标为提升用户对移动视频的主观质量评价，但优化的约束条件为有限的网络带宽、频繁波动的网络传输性能和移动终端设备处理能力、功耗等。相应的研究趋势包括：

首先，当终端用户在不同的背景环境中（如安静的图书馆或嘈杂的公交车）使用不同的移动终端设备（如大尺寸的平板电脑或相对较小屏幕的手机）观看不同形式和内容的视频时，如高分辨率低码率的屏幕视频或小分辨率大码率的教师视频，会对回放视频质量的主观评估结果存在差异。这时需要在以 PSNR 等客观评估方法的基础上，结合主观视频质量评估结果，调整优化目标。

其次，由于移动网络传输性能存在波动，使得传输协议一般需要以自适应的方式适应不同的衰落场景。因此，需要首先研究移动网络的传输特征，建立模型作为自适应算法的前置条件，同时避免调解滞后与过度调解，如自适应 FEC 编码中的编码冗余度与增益调解。

最后，需要用移动网络模拟测试平台进行算法的评估、对比与改进。开放的公共网络不适于开展理论研究，基站配置参数、并行用户的行为和信道衰落等各种因素造成无法获得精确的实验结果及结果重现。能够支持端到端的移动网络模拟测试工具，是开展移动网络下多媒体传输技术优化的重要基础。

在移动学习规模化服务方面，大规模流媒体传输技术是移动学习支持异构网络与终端的大规模移动多媒体教学场景传输的基础，但移动学习引入的一些新机制、新问题还有待进一步研究，如支持大规模异构移动学习终端、实现多源教学场景的同步传送、在低计算能力的移动学习终端上的直播与交互等。同时，如何基于虚拟化媒体云计算平台，解决大规模多媒体应用的高资源消耗特征与动态变化的服务需求，带来的供给成本与服务质量的权衡（Cost-quality trade-off）问题，提供动态服务质量可控、平台资源优化利用的大规模多媒体云服务，是基于多媒体云计算的规模化移动学习领域亟待解决的重大

挑战。

(3) 在移动学习终端的节能与资源优化利用技术领域

随着移动通信技术(3G/4G)的发展，越来越多的用户追求终端的移动性和灵活性。虽然便携式终端的电源管理技术正在不断发展，但每年电源能量密度的改善都只有5%左右，远远不能满足手机在3G/4G或者Wifi网络环境下额外的电源需求。

因此，在移动学习终端的节能与资源优化利用技术领域，考虑到移动终端的计算能力和功耗限制，需要降低算法复杂度和应用层各类反馈消息的发送频率，降低计算能耗与通信开销以延长移动终端设备的续航能力。特别地，针对多媒体视频服务的高电能消耗特征，面向无线、移动视频的节电技术研究是移动学习终端的节能与资源优化利用技术领域的发展方向。

如今，移动终端的节能技术在全世界范围内已经成为一个研究的热门方向。而移动学习终端，尤其是移动多媒体学习系统，作为一种高能源需求的应用，在电源技术发展速度无法匹配移动终端能源需求的情况下，如何系统有效地降低其能耗则是一个亟待解决的问题。相信将会有越来越多的研究者关注该领域，也将产生越来越多可以使用的研究成果。从移动学习终端的角度出发，考虑教学场景的高分辨率、低帧率特征及HTTP等流媒体载体的特殊工作机理，研究针对移动学习终端的节能与资源优化利用方法，是本领域的重要发展趋势。此外，节能与资源优化利用算法与移动学习终端设备的结合，也是本领域的一个重要发展方向。

(4) 在短视频移动学习资源的生成、转化及资源管理技术领域

移动学习资源的特殊性，是促进短视频移动学习资源的生成、转化及资源管理技术领域短视频移动学习资源生成、转换与管理进一步发展的动力。

例如，移动学习终端设备的存储资源有限，需要对已有课件资源进行压缩，而移动学习中的短视频资源的帧率一般在每秒1~5帧。研究表明，移动学习授课中，课件视频切换的频率在分钟级，基于课件视频的相关性，可以转换压缩已有的学习资源，更好地服务于移动学习应用。这是短视频学习资源转化的一个重要研究方向与内容。

此外，针对移动学习中移动终端和接入网络的异构性，移动视频学习资源需要提供不同版本的视频资源。现有的解决方案有两种：1) 存储多个版本的学习资源；2) 对视频进行实时转码。存储多个版本的学习资源需要对各类终端设备、不同网络环境客户端需要的视频格式码率等进行归类，形成多个版本视频资源，当客户端访问时根据其设备及网络状况进行服务。这种方法将增加大量的存储消耗。对视频进行实时转码的方案即根据客户端设备及网络条件分析其所需要的视频格式码率等，在服务器端实时转码为实时视频流进行服务，但此种方式会消耗较多的计算资源。因此，综合考虑移动学习环境的计算与存储资源，并利用相同移动学习资源不同版本视频资源的相关性，优化移动学习系统资源的利用，是短视频移动学习资源的生成、转化及资源管理的重要发展趋势。

(5) 在基于微课程的个性化智能移动学习技术领域

首先，微课程的生态环境建设，是基于微课程的个性化智能移动学习技术领域的一个重要研究内容与发展趋势。微课程每次说明某个知识主题的一个侧面或说清一个观点，

不仅适合于移动学习时代知识的传播，也适合学习者个性化、深度学习的需求。学习者可以随时观看微课程，掌握相关的知识点，温故而知新，对于学习者掌握和拓展相关知识很有帮助。目前，很多教学机构、商业公司投入大量人力物力推进微课的生态圈建设，并取得很大进展，未来需要共同建立大规模素材库，并支持分类及检索。从而实现资源共享，提高现有资源的利用率。

其次，基于微课的导航式学习是基于微课程的个性化智能移动学习技术领域的重要发展趋势。由于每个微课视频一般只描述特定知识主题的一个方面，因此具有不完整性。比如，对于三角形这个知识主题，有多个微课分别描述三角形的定义、三角形的各种性质或定理、三角形的应用等。不同的微视频之间存在认知依赖关系，认知依赖关系是指知识在认知方面的前序、因果、参考等关系^[98]。例如，“线性表”与“堆栈”之间存在前序关系，表明先要掌握前者，才能学习后者；而“堆栈”与“队列”之间则存在参考关系。

不同知识主题的微课混杂在一起，而面对海量微视频时人们接受信息量是有限的，这就容易造成学习者产生认知片面与偏差，进而出现信息超载的现象。最新的认知科学研究表明，认知依赖关系，特别是“因果”与“参考”两类关系，对认知具有显著影响^[99]。此外，认知依赖关系的缺失不符合激活扩散（Spreading Activation）与联想学习（Associative Learning）的认知特点^[100]。

针对微视频带来的“认知过载”问题，需要进一步研究基于认知依赖关系的可视化导航学习。其基本思路是获取微视频间的认知依赖关系，生成微课程的知识地图（Knowledge Map）。知识地图将零散的知识点或微视频以网络图的形式串起来，由浅层次向深层次递进，指出知识点学习过程中的依赖关系。

通过知识地图形成对特定知识主题各个方面的全方位展示，避免微视频的内容片面性导致的认知片面与偏差问题。可汗学院为学生提供了知识地图以及自定学习计划，让学生明确自己的学习任务，每次登录学习后学生都能在导航中看自己的学习历程以回顾自己所学的知识。基于微课的导航式学习必然成为下一代 e-Learning 及 m-Learning 的研究方向。

最后，在移动学习的个性化服务领域，目前移动学习领域尚没有建立刻画移动学习用户上下文的统一规范和标准，期待更多的研究者对这一问题进行思考和研究。

在推荐算法的研究方面，冷启动、数据稀疏和过拟合问题依然没有很好地解决。同时，目前在移动学习中如何发现情感，并且将其同显示评分、兴趣、个性、知识背景等因素转换为统一的可计算表示，进而在知识与资源推荐时能综合考虑学习者的各项非智力因素，也是挑战之一。

综上，人类学习方式越来越明显地呈现出网络化、移动化等特点，如何从人类自身的认知心理和行为特点着手开展对用户学习行为、兴趣感知的研究是本领域的发展趋势。

5.2 移动学习与 MOOC

大规模公开在线课程（MOOC）是近年来兴起的一种网络课程教学形式，它利用互

联网的相关技术，提供了基于网页的文本、图像、音频和视频等多种形式的教学课件，通过测验、论坛、在线实验等多种形式的交互方式实现教学目标，具有视频时长短、教学交互活动丰富、学习计划合理等多种特点。其中，教学视频通常采用时长较短的视频，这些视频的时长一般在3~15分钟。在视频之后，通常配有练习题或者在线实验环节用于检验学习者对视频中知识、概念的掌握情况。通过这种短视频，可以充分吸引学生在线学习时的注意力，避免由于长时间观看引起的注意力不集中等问题。

从2012年起，众多世界一流的大学陆续在网络学习平台上提供该校精品免费课程，随着Coursera、Udacity、edX三大课程提供商的兴起，更多的学校加入MOOC网络学习平台，旨在为学生提供系统学习的机会，因此2012年也被人们称为“MOOC元年”。据2013年Jordan的统计数据显示，包括Coursera、Udacity、edX、NAMoodle、Canvas.net及Class2go在内的几大主要MOOC平台的单门课程注册人数就已到达5万人次，其中某些经典课程的最高注册人数更是达到23万人次^[101]。如今，随着MOOC的不断普及，这些平台对注册用户数均以十万为单位进行计数。

目前，国内的MOOC平台最具代表性的是学堂在线。国外的MOOC平台主要如下表所示。

名称/网址	类型	参与学校
edX https://www.edx.org/	非盈利	MIT Harvard University UC Berkeley Kyoto University Australian National University University of Queensland
Coursera/ https://www.coursera.org/	商业	University of Maryland Wharton School University of Virginia Stanford University University of Tokyo
Udacity/ http://www.udacity.com/	商业	Georgia Institute of Technology
iVersity/ https://iversity.org/	非盈利	Universidad Autonoma de Madrid University of Florence University of Hamburg
FranceUniversitéNumérique/ http://www.france-universite-numerique.fr/	非盈利	Conservatoire National des Arts et Métiers École normale supérieure de Cachan University of Paris-Sud
Eliademy/ http://eliademy.com/	商业	Aalto University Executive Education
CanvasNetwork/ http://www.instructure.com/	商业	Santa Clara University University of Utah Université Lille1
OpenLearning/ https://openlearning.com/	商业	University of New South Wales Taylor's University University of Canberra

(续)

名称/网址	类型	参与学校
AcademicEarth/ http://academicearth.org/	非盈利	UC Berkeley UCLA University of Michigan Oxford University
FutureLearn/ http://futurelearn.com/	非盈利	University of Reading Open University Monash University Trinity College Dublin Warwick University University of Bath
Acade.me/ http://acade.me/	商业	Universidad Latina
MOOEC/ http://www.mooec.com/	非盈利	University of Queensland Griffith University University of Technology
Novoed/ http://novoed.com/	商业	Stanford University Carnegie Foundation Universidad Católica de Chile
WizIQ/ http://www.wiziq.com/	商业	IITDelhi Des Moines Area Community College
KhanAcademy/ https://www.khanacademy.org/	非盈利	无
Saylor.org/ http://www.saylor.org/	非盈利	无
Udemy/ https://www.udemy.com/	商业	无

目前，移动设备已成为进入网络的最主要方式，若少了移动学习，MOOC 必将受到极大限制，因此移动学习与 MOOC 之间必然存在紧密的交集。2007 年，Kukulska-Hulme 和 Traxler 二人就曾表示：未来如果能够将移动技术用于支持“非正式的、情境性、个性化”的移动学习，那么这种具有创新性和挑战性的学习形式将会更加令人期待^[102]。MOOC 的前两个字母 M 与 O 分别被定义为“开放性”和“在线性”，再加上 MOOC 平台本身具有的自主性，完美地诠释了 Kukulska-Hulme 和 Traxler 提出的“非正式的、情境性、个性化”的移动学习。同年，Winsters 曾将移动学习的特点描述为以下三方面：移动学习可以帮助学习者在不同的情境中构建属于其自己的知识体系；移动学习可以帮助学习者建构理解；移动学习可以帮助学习者提供更加灵活的时空环境^[103]。实际上，基于 MOOC 的学习同样具有上述特点，MOOC 不仅打破了传统学习对于时间与空间的限制，还能够使学生获得合理的自主学习计划和系统的理论知识体系，它将所有学习资源集中在云端，允许那些有意愿的学习者随时随地地学习。因此，移动学习与 MOOC 之间实际上是密切相关的，二者互相影响、彼此依赖。

可见，移动学习与 MOOC 的结合具有重大的意义，它是网络学习时代酝酿出的一种

全新的个性化学习模式。这种学习模式是一个理想的非正式学习状态，它的移动特性可以基本满足处于动态中的学习者的需求，它的 MOOC 特性可以为学习者充分利用零碎时间进行学习提供便利，真正意义上实现了人类按需学习的理想。基于 MOOC 的移动学习模式可以通过移动设备随时随地进行网络学习，它是移动计算技术、无线通信技术和计算机科学技术等多种技术相结合的产物，在很大程度上拓展了学习场所与时间的范围，并提高了学习效率。

如今，移动学习技术与 MOOC 仍在迅猛地发展，并且两者间存在着众多共性，加上移动设备将会越来越灵活，大多数基于互联网的应用都可以通过移动设备实现，必将开启一个移动学习技术与 MOOC 相融合的发展模式。当然，任何模式的发展必将具有其自身的限制。首先，伴随着移动学习与 MOOC 的结合，我们必须清楚其对使用资源的限制与消耗；其次，由于移动互联网的接入方式会使用户产生大量的花费，必须考虑到数据传输过程中所消耗的费用；最后，随着不同类型移动设备的大量涌现，必须解决其在各类系统间不兼容的情况，实现移动平台的无缝连接。

尽管这种学习模式的发展仍然存在技术问题、费用问题、社会问题等众多问题，但是目前这些问题都已经有了较为明确的解决方案。移动学习是 MOOC 与大量学习群体能够保持联系的理想方式，这种结合比传统学习形式更能促进学生高质量的学习，更重要的是它打破了传统学习的时空限制，使得学习形式变得更加方便、直接和灵活。所以说，这两种学习形式的结合仍是大势所趋，其发展具有巨大的潜力，未来的前景也将更加光明。这种全新的学习模式是移动通信、网络技术与当代教育有机结合的结果，也是现代教育技术的前沿成果，其推广和发展必将实现对教育资源的充分利用，并极大程度地提高人们的学习效率。随着我们对移动学习与 MOOC 不断深入的认识与实践，这种全新的学习模式必将打破对时间、空间和地域的限制，为教育模式及理念的变革带来巨大的发展空间，展现出其独特魅力。

5.3 移动学习与大数据分析

本节将介绍移动学习与大数据分析技术的相互影响与作用，展望基于大数据分析的移动学习个性化、智能化的发展趋势；介绍移动学习大数据分析所需的存储与计算技术发展趋势。

（1）大数据分析的移动学习个性化、智能化的发展趋势

从国家教育战略的角度讲，衡量一个国家教育的成功与失败必须以其加强自身科学技术及效率的进步能力为准则^[104]。据此，美国白宫在 2012 年发布的科学技术政策报告^[105]中强调：在应对同时代的中、俄、印、巴的强劲崛起，使得美国把大数据及其相关技术看做“解决国家最大压力挑战的一些技术之一”，同时“通过改进它的能力来从巨大复杂的数字化数据中抽取知识和观点。”与此同时，联合国教科文组织和美国教育信息化协会认为，未来的移动学习在满足了大幅增加的职业培训和终生学习的多元化渠道选择需求的同时，也面临着学习资源自身的数据增长和学习者访问不同网

站的不同资源而产生的各种类型数据增长压力。因此，移动学习和大数据技术将相互影响与作用。

首先，从时空演化的角度看，移动学习为大数据技术提供数据基础与需求。①移动学习的内容，即知识本身，存在碎片化、分布广、知识密度稀疏化，且类型多样（Various）、容量（Volume）大且动态变化速度（Velocity）迅速。毫无疑问，这些会对学习者的认知带来巨大的挑战，“如何利用大数据分析技术合适的组织知识、检索知识、利用知识等”问题呼之欲出。②学习者自身产生的数据多样化、复杂化。学习者的个人信息、评估信息、GPS 坐标、完成任务或作业的时间、所产生包括文本、视音频等媒体信息等，通过日积月累将形成巨量的“大数据”。以上对大数据中的采集、综合处理、存储、搜索、共享、转换、分析与可视化提出新的挑战。

其次，从相互作用的角度看，移动学习也为大数据技术的应用提供了载体和平台。学习分析（Learning Analytics）被认为是未来 50 年大数据技术在移动学习中的重要应用^[106]。学习分析是从教育的大数据或者与学生相关的海量数据中辨别他们的学习行为发展趋势和模式，以此促进个性化的高等教育支持系统。学习分析的目标是^[107]：①学习流程优化。颠覆传统课堂的界限，推动混合学习（Blended Learning）和协同学习（Collaborative Learning）。②决定移动学习中的场景选择和协同交互模式选择问题。从大数据视角的行为科学角度分析，不仅在空间上，而且在时间上都能够充分理解学习者的学习行为。通过长时间的分析和洞察学习行为的模式，发现他们在什么样的场景下更倾向于做什么。③教师教学指导与培训^[106,108]。通过分析学习者的喜好、收集各种评估，可以让教师实时掌握学生动态，了解教学效果，及时调整教学策略。④人们日益关注利用新的数据源实现个性化的学习体验和绩效评估。

最后，当前知识碎片化和移动学习时代，大数据与移动学习相结合的机遇与挑战共存。其发展趋势为：①个性化学习推荐。如何面对碎片化知识，发现并推荐给学习者合适的知识；②突破现有的学习理论。现行教育与教学的固有理念与实践模式成为新技术得到广泛采用的滞障。而如何利用碎片化的短时间来学习成体系、成系统的知识章节成为新学习理论形成的一个重要衡量指标；③协同学习模式发现与分析；④大数据中数据收集、分析、共享与个人隐私的冲突急需解决。

（2）移动学习大数据分析所需的存储与计算技术发展趋势

随着移动学习应用的不断发展，其应用的情况也会变得更加复杂。数据的来源多样，学习者的特征也更加多样，为各种不同的学习者提供有价值的信息，评价移动学习的效果，提高移动学习者的学习效率，研究移动学习与传统学习之间的区别特点，挖掘学习者的个性特点等各个方面都需要更加深入且高效的技术手段提供支持，所以利用大数据的技术、工具和方法，能够对移动学习的数据分析提供进一步研究发展的基础，满足科研与应用的需求。

移动学习终端的增加以及信息采集需求的增加，将会直接引起原始采集数据量的显著攀升，采用大规模的分布式存储系统可以保存这些海量的原始采集数据，成为必不可

少的基础平台。利用分布式存储系统的横向扩展特性，可以为原始数据的保存带来稳定、可靠、能够扩容的存储空间，确保原始数据能够持久记录，并且可以被并行计算框架所利用，提高各种分析计算的效率。

移动学习涉及的知识资源包含多种形式、内容以及关联关系的数据，除了基本的文件存储以外，可以利用大规模的图数据库对这些具有大量复杂关联关系的数据建立模型，构建知识地图的复杂网络，利用对复杂网络分析研究的方法，研究移动学习资源的内容，为学习者的移动学习推荐提供基础支持。

由于数据的维度不断攀升，会有大量的潜在信息隐式地蕴含在数据的维度之中，应用大数据分析方法中的深度学习方法，能够对这些维度中的信息进行抽取，发现数据内在的意义，这对于研究移动学习者的学习特性、移动学习环境下的教学方法与教学工具设计、移动学习效果的评价等方面都有重要作用。

对于海量的移动学习记录信息，其数据分析的方法也将从单机处理程序变为基于并行化计算引擎的分析方法，利用 MapReduce 提供的并行计算框架，对分布式存储系统中的数据进行分析，提高对海量数据的处理效率。

面向解决移动学习中情感、个性等复杂计算场景要求的迭代过程，需要在基本的分布式文件存储系统之上，采用 Spark 等能够支持并行的机器学习计算框架，高速读写数据的分布式内存存储系统和 NoSQL 存储系统，充分利用计算集群的内存资源，提高迭代计算的数据传输速度。

为了更加快速地分析移动学习者的需求、预测移动学习者的行为、提供有价值的参考信息，需要进一步地将流式实时计算应用在移动学习的整体架构设计中，采用增量计算的方法，对实时到达的数据进行处理和计算，减少批量处理过程的时间消耗。

上述的关键技术与工具方法，可以有效地解决移动学习相关的大数据问题，为更好地提供研究与应用提供基础的保障，这是移动学习在大数据时代的研究发展方向。

6 结束语

综上，移动学习旨在构建具有高度移动性、个性化、智能化、协作性的网络认知环境，涉及计算机科学、认知科学、教育学等多门学科。因此，移动学习作为一项典型而广泛的移动互联网应用，如何真正满足人们的移动化认知学习需求，依然是一个富有挑战性的研究内容。如何提供大规模、高质量、个性化、智能化的移动学习服务，依然是当前移动学习亟待解决的核心问题。

与认知科学、教育学领域众多移动学习的研究分析成果不同，本报告着重从计算机科学的领域出发，以移动计算、云计算、网络多媒体、知识获取与个性化服务等相关技术为基础，概要介绍了移动学习技术的研究进展与趋势。通过分析比较移动学习技术的国际研究现状与国内研究进展，并对移动学习的发展趋势进行总结与展望，希望能使阅

读者深层次地了解人人皆可、自主学习、天地结合、无处不网、各类终端、实时在线的泛在移动学习技术。

基于移动学习的研究进展与趋势的分析以及团队在 e-Learning/m-Learning 领域的长期实践，从重大应用中探索移动学习应用基础理论与关键技术，并回归实际 e-Learning/m-Learning 应用是移动学习技术研究与实践的必由之路，笔者期望移动学习的研究者与实践者一起，致力于提高国产软件的自主发展能力，大力推动业务创新和服务模式创新，强化信息技术在 e-Learning/m-Learning 领域的运用，获得自主知识产权，从而在新一代信息技术领域占领制高点，进而提升我国在此领域的竞争力。在移动互联网、MOOC、大数据分析等技术的引领下，走出一条具有自身特色的远程教育之路。

参考文献

- [1] H Crompton. Handbook of mobile learning[M]. Florence, KY: Routledge. 2013 : 3-14.
- [2] M-learning[EB/OL]. <http://en.wikipedia.org/wiki/MLearning>.
- [3] W Zhu, C Luo, J Wang, S Li. Multimedia Cloud Computing[J]. IEEE Signal Processing Magazine, 2010, 28 : 59-69.
- [4] X Wen, K Chen, Y Chen, et al. Virtualknotter: Online Virtual Machine Shuffling for Congestion Resolving in Virtualized Datacenter [C]. Proceedings of the 32nd IEEE International Conference on Distributed Computing Systems (ICDCS), Macau, China, 2012 : 12-21.
- [5] R Jeyarani, N Nagaveni, R Vasanth Ram. Self Adaptive Particle Swarm Optimization for Efficient Virtual Machine Provisioning in Cloud [J]. International Journal of Intelligent Information Technologies (IJIIT), 2011 , 7(2) : 25-44.
- [6] J Xu, J Fortes. Multi-objective virtual machine placement in virtualized data center environments [C]. Proceedings of the IEEE/ACM International Conference on Green Computing and Communications & 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing, Hangzhou, China, 2010 : 179-188.
- [7] A Verma, G Dasgupta, T Nayak, et al. Server Workload Analysis for Power Minimization Using Consolidation [C]. Proceedings of the 2009 conference on USENIX Annual technical conference, San Diego, USA, 2009 : 28-28.
- [8] P Padala, K G Shin, X Zhu, et al. Adaptive Control of Virtualized Resources in Utility Computing Environments [C]. Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007, New York, USA, 2007 : 289-302.
- [9] X Liu, X Zhu, P Padala, et al. Optimal Multivariate Control for Differentiated Services on a Shared Hosting Platform[C]. Proceedings of the 46th IEEE Conference on Decision and Control, New Orleans, LA, USA, 2007 : 3792-3799.
- [10] J Heo, X Zhu, P Padala, et al.. Memory Overbooking and Dynamic Control of Xen Virtual Machines in Consolidated Environments [C]. Proceedings of the IFIP/IEEE International Symposium on Integrated

- Network Management (IM 2009), New York, USA, 2009: 630-637.
- [11] M A Kjær, M Kihl, A Robertsson. Resource Allocation and Disturbance Rejection in Web Servers using Slas and Virtualized Servers [J]. IEEE Transactions on Network and Service Management, 2009, 6(4): 226-239.
- [12] D Ardagna, R Mirandola, M Trubian, et al. Run-time Resource Management in SOA Virtualized Environments[C]. Proceedings of the 1st International Workshop on the Quality of Service-Oriented Software Systems (QUASOSS 09), Amsterdam, Netherlands, 2009: 39-46.
- [13] J Xu, M Zhao, J Fortes, et al. Autonomic Resource Management in Virtualized Data Centers Using Fuzzy Logic-based Approaches[J]. Cluster Computing, 2008, 11(3): .213-227.
- [14] Y Wu, C Wu, B Li, et al. Cloudmedia: When Cloud on Demand Meets Video on Demand[C]. Proceedings of the 31st International Conference on Distributed Computing Systems (ICDCS), Minneapolis, USA, 2011: 268-277.
- [15] D Niu, H Xu, B Li, et al. Quality-assured Cloud Bandwidth Auto-scaling for Video-on-demand Applications [C]. Proceedings of the IEEE INFOCOM Conference, Orlando, USA, 2012: 460-468.
- [16] S Pawar, S El Rouayheb, H. Zhang, et al. Codes for a Distributed Caching Based Video-on-demand System [C]. Proceedings of the 2011 45th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, USA, 2011: 1783-1787.
- [17] W Zhang, Z Mo, C Chen, et al. CBC: Caching for Cloud-based VOD Systems[J]. Multimedia Tools and Applications, 2013: 1-24.
- [18] M Kim, Y Cui, S Han, et al. Towards Efficient Design and Implementation of a Hadoop-based Distributed Video Transcoding System in Cloud Computing Environment [J]. International Journal of Multimedia & Ubiquitous Engineering, 2013, 8(2): 213-224.
- [19] Z Huang, C Mei, L Li, et al. Cloudstream: Delivering high-quality streaming videos through a cloud-based svc proxy[C]. Proceedings of the IEEE INFOCOM Conference, Shanghai, China, 2011: 201-205.
- [20] S Lin, X Zhang, Q Yu, et al. Parallelizing Video Transcoding with Load Balancing on Cloud Computing [C]. Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, China, 2013: 2864-2867.
- [21] H Wang, Y Shen, L Wang, et al. Large-scale Multimedia Data Mining Using MapReduce Framework[C]. Proceedings of the 2012 IEEE 4th International Conference on Cloud Computing Technology and Science (CloudCom), Taipei, Taiwan, 2012: 287-292.
- [22] K Liu, T Zhang, L Wang. A New Parallel Video Understanding and Retrieval System[C]. Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Suntec City, Singapore, 2010: 679-684.
- [23] H-Y Huang, T-C Huang, N Chilamkurti, et al.. Adaptive Forward Error Correction With Cognitive Technology Mechanism for Video Streaming over Wireless Networks[C]. Computer Communication Control and Automation (3CA), 2010 International Symposium on, IEEE: 2010: 519-521.
- [24] J Xiao, T Tillo, C Lin, et al.. Dynamic Sub-GOP Forward Error Correction Code for Real-time Video Applications[J]. Multimedia, IEEE Transactions on, 2012, 14(4), 1298-1308.
- [25] A Majumda, DG Sachs, IV Kozintsev, et al.. Multicast and Unicast Real-time Video Streaming over Wireless LANs[J]. Circuits and Systems for Video Technology, IEEE Transactions on, 2002, 12(6), 524-534.

- [26] Z Rongfu, Z Yuanhua, H Xiaodong. Content-adaptive Spatial Error Concealment for Video Communication [J]. Consumer Electronics, IEEE Transactions on, 2004, 50(1), 335-341.
- [27] Q Peng, T Yang, C Zhu. Block-based Temporal Error Concealment for Video Packet Using Motion Vector Extrapolation [C]. Communications, Circuits and Systems and West Sino Expositions, IEEE 2002 International Conference on, IEEE: 2002; 10-14.
- [28] C Mehlührer, M Wrulich, JC Ikuno, et al. . Simulating the Long Term Evolution Physical Layer[C]. Proc. of the 17th European Signal Processing Conference (EUSIPCO 2009) , Glasgow, Scotland, 2009; 124.
- [29] X Xu, F Schroeder, B Gevrekce, et al. . A Physical Layer Simulator Based on Radio Wave Propagation for LTE Cellular Networks [C]. Antennas and Propagation (EuCAP) , 2013 7th European Conference on, IEEE: 2013; 1007-1010.
- [30] M Mezzavilla, M Miozzo, M Rossi, et al. . A Lightweight and Accurate Link Abstraction Model for The Simulation of LTE Networks in ns-3 [C]. Proceedings of the 15th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, ACM: 2012; 55-60.
- [31] F Guidolin, L Badia, M Zorzi, Implementation of 2×2 MIMO in an LTE Module for the ns3 Simulator[C]. Computer Aided Modeling and Design of Communication Links and Networks (CAMAD) , 2012 IEEE 17th International Workshop on, IEEE: 2012; 281-285.
- [32] N Baldo, M Miozzo, M Requena-Esteso, et al. . An Open Source Product-oriented LTE Network Simulator Based on ns-3 [C]. Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems, ACM: 2011; 293-298.
- [33] P Munoz, I de la Bandera, F Ruiz, et al. . Developing a Computationally-Efficient Dynamic System-Level LTE Simulator[J]. 4G Wireless Communication Networks: Design Planning and Applications, 2013, 265.
- [34] N Baldo, M Requena-Esteso, M Miozzo, et al. . An Open Source Model for the Simulation of LTE Handover Scenarios and Algorithms in ns-3[C]. Proceedings of the 16th ACM international conference on Modeling, analysis & simulation of wireless and mobile systems, ACM: 2013; 289-298.
- [35] W Zhang, Q Zheng. Multi-Channel Live Streaming in Service Overlay Network [J]. Multimedia Tools and Applications, Springer, to appear, <http://www.springerlink.com/content/f5887k0016151817/>.
- [36] W Zhang, Q Zheng, H. Li. An Overlay Multicast Protocol for Live Streaming and Delay-guaranteed Interactive Media [J]. Journal of Network and Computer Applications, Elsevier, to appear, <http://dx.doi.org/10.1016/j.jnca.2011.02.013>.
- [37] W zhang, Z li, Q zheng. SAMP Supporting Multi-source Heterogeneity in Mobile P2P IPTV System[J]. IEEE Transactions on Consumer Electronics, 2013, 59(4): 772-778.
- [38] I Trajkovska, J Salvachua, A M Velasco. A Novel P2P and Cloud Computing Hybrid Architecture for Multimedia Streaming with QoS Cost Functions [C]. Proceedings of ACM Multimedia, 2010; 1227-1230.
- [39] B Zhao, Q Zheng. Energy-Aware Web Browsing in 3G Based Smartphones[C]. Proceeding of IEEE 33rd International Conference on Distributed Computing Systems, IEEE: Philadelphia, PA, 2013; 165-175.
- [40] B Zhao, B C Tak, G Cao. Reducing the Delay and Power Consumption of Web Browsing on Smartphones in 3G Networks [C]. Proceeding of 31st International Conference on Distributed Computing Systems, Minneapolis, MN, 2011; 413-422.
- [41] M Siekkinen, M Ashraful, M Aalto. Streaming over 3G and LTE: How to Save Smartphone Energy in Radio Access Network-friendly Way[C]. Proceeding of ACM Workshop on Mobile Video, ACM: Oslo, Norway, 2013; 13-18.

- [42] X Lu, E Erkip, Y Wang, et al. Power Efficient Multimedia Communication over Wireless Channels [J]. Selected Areas in Communications, 2003, 21(10): 1738-1751.
- [43] B Anand, A Ananda, M Chan, et al. Game Action Based Power Management for Multiplayer Online Game [C]. Proceeding of the 1st ACM workshop on Net-working, systems, and applications for mobile handhelds. New York, USA, 2009: 55-60.
- [44] Guo P J, Kim J, Rubin R. How Video Production Affects Student Engagement: An Empirical Study of Mooc Videos [C]. Proceedings of the first ACM conference on Learning @ scale conference. ACM: Atlanta, Georgia, USA. 2014: 41-50.
- [45] 于青青, 冯雪松. 基于内录式的 MOOCs 视频制作与分析[J]. 中国教育信息化: 基础教育, 2014 (1): 14-16.
- [46] Chuncong Xu, Xiaomeng Huang, Nuo Wu, et al. Using Memcached to Promote Read Throughput in Massive Small-File Storage System [C]. Proceedings of IEEE 9th International Conference on Grid and Cooperative Computing, Nanjing, China, 2010: 24-29.
- [47] L Morris. Little Lectures? [J]. Innovative Higher Education, 2009, 34(2), 67-68.
- [48] D Edge, E Searle, K Chiu, et al. MicroMandarin: Mobile Language Learning in Context [C]. Proceeding of CHI 2011, Vancouver, BC, Canada, 2012: 3169-3178.
- [49] J A Russell. A Circumplex Model of Affect [J]. Personality and Social Psychology, 1980, 39: 1161-1178.
- [50] A Ortony, G Clore, A Collins. The Cognitive Structure of Emotions [M]. Cambridge University, 1988.
- [51] 肖觅, 孟祥武, 史艳翠. 一种基于移动用户行为的回路融合社区发现算法 [J]. 电子与信息学报, 2012. 34(10): 2369-2374.
- [52] K Verbert, N Manouselis, X Ochoa, et al. Context-Aware Recommender Systems for Learning A Survey and Future Challenges [J]. IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, 2012, 5(4): 318-335.
- [53] P Dolog and W Nejdl. Challenges and Benefits of the Semantic Web for User Modelling [C]. Proc. Workshop Adaptive Hypermedia and Adaptive Web-Based Systems (AH '03), 2003: 99-112.
- [54] S Shishehchi, S Y Banihashem, N A M Zin, et al. Review of Personalized Recommendation Techniques for Learners in E-learning Systems [C]. 2011 International Conference on Semantic Technology and Information Retrieval, IEEE: Putrajaya, Malaysia, 2011: 277-281.
- [55] D Frohberg. Mobile Learning is Coming of Age-What we have and what we still miss [C]. Proceedings of DeLF 2006 e-Learning Fachtagung Informatik, Darmstadt, Germany. 2006: 327-338.
- [56] 杨俊峰, 王以宁. 移动教育比较研究 [N]. 中国计算机报, 2006-04-24B13.
- [57] M-Learning Project. mediaBoard [EB/OL]. <http://www.m-learning.org/products/mediaboard.htm>.
- [58] G Vavoula, J Meek, M Sharples, et al. A Lifecycle Approach to Evaluating MyArtSpace [C]. Proceedings of the 4th IEEE International Workshop on Wireless, Mobile and Ubiquitous Technology in Education (WMTE '06), Athens, Greece. 2006: 18-22.
- [59] Ambient Insight [EB/OL]. <http://www.ambientinsight.com/Reports/MobileLearning.aspx#section4>.
- [60] ASTD2012 培训行业研究报告分享 [EB/OL]. <http://www.online-edu.org/html/2013/17234.html>.
- [61] mobile-learning-tools [EB/OL]. <http://elearningindustry.com/mobile-learning-tools-mlearning-edtech>.
- [62] 杜海鹏, 张未展, 郑庆华. 大规模多元化移动式学习技术 [J]. 中国科技论文在线, 2011, 6(10): 761-764.
- [63] Kong X, Lin C, Jiang Y, et al. Efficient Dynamic Task Scheduling in Virtualized Data Centers with Fuzzy Prediction [J]. Journal of network and Computer Applications, 2011, 34(4): 1068-1077.

- [64] Song Y, Wang H, Li Y, et al. Multi-tiered On-demand Resource Scheduling for VM-based Data Center [C]. Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid. IEEE Computer Society, 2009: 148-155.
- [65] 李建锋, 彭舰. 云计算环境下基于改进遗传算法的任务调度算法[J]. 计算机应用, 2011, 31(1): 184-186.
- [66] 华夏渝, 郑骏, 胡文心. 基于云计算环境的蚁群优化计算资源分配算法[J]. 华东师范大学学报: 自然科学版, 2010, 1(1): 127-134.
- [67] 孙大为, 常桂然, 李凤云, 等. 一种基于免疫克隆的偏好多维 QoS 云资源调度优化算法[J]. 电子学报, 2011, 39(8): 1824-1831.
- [68] Wei G, Vasilakos A V, Zheng Y, et al. A Game-theoretic Method of Fair Resource Allocation for Cloud Computing Services[J]. The Journal of Supercomputing, 2010, 54(2): 252-269.
- [69] 葛新, 陈华平, 杜冰, 等. 基于云计算集群扩展中的调度策略研究[J]. 计算机应用研究, 2011, 28(3).
- [70] 尹红军, 李京, 宋浒, 等. 云计算中运营商效益最优的资源分配机制[J]. 华中科技大学学报(自然科学版), 2011, 1.
- [71] You X, Xu X, Wan J, et al. Ras-m: Resource Allocation Strategy Based on Market Mechanism in Cloud Computing[C]. ChinaGrid Annual Conference, 2009. ChinaGrid'09. Fourth. IEEE, 2009: 256-263.
- [72] 苏放, 甄雁翔, 景晓军. 模糊综合评判的融合网络 2 种丢包原因区分[J]. 北京邮电大学学报, 2009, 32(3): 60-4.
- [73] 林晓峰. 自适应 FEC 丢包恢复技术的研究[D]. 北京邮电大学, 2010.
- [74] 缪西梅. 无线网络中支持 H.264 的增强前向纠错算法[J]. 计算机应用, 2008, 28(9): 2225-9.
- [75] 秦艳辉. 保证视频通信质量的选择重传技术研究与实现[D]. 西安电子科技大学, 2010.
- [76] 李子诺. 基于视频内容重要性的选择重传方法研究[D]. 西安电子科技大学, 2010.
- [77] 李强, 何骥鸣, 明艳. 基于边缘检测及方向加权的 H.264 帧内错误隐藏算法[J]. 计算机应用研究, 2010, 27(12): 4798-800.
- [78] 熊春彬, 张有志, 李庆涛, 等. 基于自适应叠加的 H.264 时域错误隐藏算法[J]. 计算机工程, 2010, 36(12): 236-7, 47.
- [79] 张瑞, 洪佩琳, 李津生, 等. 无线网络中一种改进的 TCP 拥塞控制机制[J]. 电路与系统学报, 2006, 06): 7-13.
- [80] 肖甫, 王汝传, 孙力娟, 等. 基于 TCP 友好的无线网络拥塞控制机制研究[J]. 计算机科学, 2010, 07): 50-3.
- [81] 唐辉, 张国杰, 黄建华, 等. 一种混合 P2P 网络模型研究与设计[J]. 计算机应用, 2005, 03): 521-4 +35.
- [82] 孙辉, 张晋豫. 基于推拉结合机制的 P2P 流媒体分发算法[J]. 软件学报, 2013, 05: 43-7.
- [83] 吕广娜. 基于智能终端多媒体业务云计算平台中关键技术的研究[D]. 北京邮电大学, 2013.
- [84] X Li, M Dong, Z Ma. Felix Fernandes. GreenTube: Power Optimization for Mobile Video Streaming via Dynamic Cache Management[C]. Proceeding of MM '12, Nara, Japan, 2012: 279-288.
- [85] 张宇, 刘新, 叶德建. 基于分布式流媒体计算框架的转码系统的设计与实现[J]. 计算机应用与软件, 2013.30(9): 92-95.
- [86] 赵跃龙, 谢晓玲, 蔡咏才, 等. 一种性能优化的小文件存储访问策略的研究[J]. 计算机研究与发展, 2012, 49(7): 1579-1586.

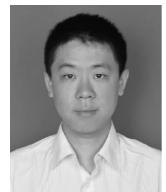
- [87] 刘小俊, 徐正全, 潘少明. 一种结合 RDBMS 和 Hadoop 的海量小文件存储方法[J]. 武汉大学学报(信息科学版), 2013, 38(1): 113-115.
- [88] 刘高军, 王帝澳. 基于 Redis 的海量小文件分布式存储方法研究[J]. 计算机工程与科学, 2013, 35(10): 58-64.
- [89] 许春聪, 黄小猛, 徐鹏志, 等. CarrierFS: 基于虚拟内存的分布式文件系统[J]. 华中科技大学学报(自然科学版), 2010, 38(Suppl. iv): 37-42.
- [90] 陈卓, 熊劲, 马灿. 基于 SSD 的机群文件系统元数据存储系统[J]. 计算机研究与发展, 2012, 49(Suppl.): 269-275.
- [91] 孟祥武, 王凡, 史艳翠, 等. 移动用户需求获取技术及其应用[J]. 软件学报, 2014, 25(03): 439-456.
- [92] 王立才, 孟祥武, 张玉洁. 上下文感知推荐系统[J]. 软件学报, 2012, 23(01): 1-20.
- [93] 孟祥武, 胡勋, 王立才, 等. 移动推荐系统及其应用[J]. 软件学报, 2013, 24(01): p. 91-108.
- [94] F Tian, P Gao, L Li, et al. Recognizing and Regulating E-learners' Emotions Based on Interactive Chinese Texts in E-learning Systems[J]. Knowledge-Based Systems, 2014, 55: 148-164.
- [95] 邹博伟, 张宇, 范基礼, 等. 基于改进 TextTiling 方法的用户新兴趣发现的研究[J]. 计算机研究与发展, 2009, 46(9): 1594-1600.
- [96] 王立才, 孟祥武, 张玉洁. 移动网络服务中基于认知心理学的用户偏好提取方法[J]. 电子学报, 2011, 39(11): 2547-2553.
- [97] 周涛, 韩筱璞, 闫小勇, 等. 人类行为时空特性的统计力学[J]. 电子科技大学学报, 2013, 42(4): 481-540.
- [98] J Liu, L Jiang, ZH Wu, et al. Mining Learning-Dependency Between Knowledge Units From Text[J]. The VLDB Journal, 2011, 20(3): 335-345.
- [99] Pvd Broek. Using Texts in Science Education: Cognitive Processes and Knowledge Representation [J]. Science, 2010, 328(5977): 453-456.
- [100] J Cortese. Internet Learning and the Building of Knowledge2007, Youngstown: Cambria Press.
- [101] JordanK. MOOC Completion Rates: TheData [EB/OL]. <http://www.katyjordan.com/MOOCproject.html>, 2013-09-22.
- [102] H Beetham, R Sharpe. Rethinking Pedagogy for a Digital Age: Designing for 21st Century Learning[M]. NY, USA: Routledge, 2013: 352.
- [103] M Sharples. What is Mobile Learning[R]. Big issues in mobile learning Winters, 2006: 5.
- [104] Tyack D, Cuban L. Progress or Regress? In L. Iura (Ed.), The Jossey-Bass reader on school reform (pp. 5-42). 2001, San Francisco, CA: Wiley Company.
- [105] Office of Science and Technology Policy. (2012). Obama administration unveils "Big Data" initiative; Announces \$200 million in new R&D investments. Retrieved from: http://www.whitehouse.gov/.../big_data_press_release_final_2.pdf.
- [106] United Nations Educational, Scientific and Cultural Organization (UNESCO). The Future of Mobile Learning: Implications for Policy Makers and Planners, 2013. (URL: unesdoc.unesco.org/images/0021/002196/219637e.pdf).
- [107] LAK. 2011. 1st International Conference on Learning Analytics and Knowledge 2011. Banff, Alta, Author. <https://tekri.athabascau.ca/analytics/about>.
- [108] 美国新媒体联盟(New Media Consortium, NMC)和美国高校教育信息化协会. 新媒体联盟地平线报告: 2013 高等教育版[J]. 2013, 北京开放大学, 中国.

作者简介

郑庆华 男，博士，西安交通大学电信学院计算机系教授，CCF 高级会员，IEEE/ACM 会员，国家杰出青年基金获得者，教育部长江学者特聘教授，国家“万人计划”科技创新领军人才，国家“新世纪百千万人才工程”。主要从事智能 e-Learning 理论及技术、网络舆情、可信软件等方向研究。先后主持承担 NSF 重点项目、863 计划、国家科技支撑计划、核高基重大专项等课题，研究成果获得 2 项国家科技进步二等奖、5 项省部级科技进步一等奖，以及国家教学成果一等奖和二等奖。获得国家发明专利授权 24 项，发表论文 160 余篇，SCI、EI 收录 150 余篇。曾获中国科协“求是”杰出青年奖、中国青年科技奖、宝钢优秀教师特等奖等荣誉。



张未展 男，博士，西安交通大学电信学院计算机系副教授，硕士生导师。CCF 高级会员，IEEE 会员，ACM 会员。主要研究方向为面向 e-Learning 的网络多媒体技术，涉及移动网络、云计算等相关研究领域。近年来，主持和参与国家自然科学基金 2 项，核高基重大专项课题 1 项，国家科技支撑计划课题 3 项，发改委 CNGI 专项 2 项。在相关国际期刊、会议上发表论文 20 余篇，申请、授权发明专利 10 项。获中国电子学会科技进步 1 等奖。



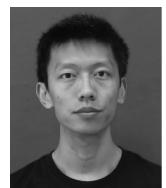
田 锋 博士，西安交通大学电信学院自动化系副教授，博士生导师。IEEE 会员。主要研究方向为智能网络学习环境与个性化服务技术，涉及数据挖掘、情感计算等相关领域。近年来主持和参与国家自然科学基金 5 项，国家科技支撑课题 3 项。在相关国际期刊、会议发表论文 50 余篇，申请、授权发明专利 7 项。



魏笔凡 博士，西安交通大学网络教育学院工程师。主要研究方向为 Web 信息抽取，Web 文本挖掘。近年来参与国家科技支撑计划课题、国家自然科学基金等 3 项。在 IEEE TKDE、Journal of Web Engineering、WWW、ICONIP 国际期刊、会议上发表多篇论文，申请、授权发明专利 2 项。



杜海鹏 硕士，西安交通大学网络教育学院工程师。主要研究方向为面向 e-Learning 的网络多媒体技术，涉及移动网络传输优化等相关研究领域。近年来，参与国家自然科学基金 1 项，核高基重大专项课题 1 项，国家科技支撑计划课题 3 项。



社会媒体搜索技术：挑战及进展

CCF 办公自动化专业委员会

于戈¹ 王大玲¹ 申德容¹ 冯时¹ 李瑞轩² 李玉华² 汤庸³ 邢春晓⁴ 于旭⁵ 姜安琦⁶

¹东北大学信息科学与工程学院，沈阳

²华中科技大学计算机科学与技术学院，武汉

³中山大学计算机学院，广州

⁴清华大学计算机科学与技术系，北京

⁵香港中文大学系统工程与工程管理系，香港

⁶新浪门户广告产品技术部，北京

摘要

随着 Web 2.0 时代的到来，各种社会媒体（如 QQ、微博、微信等）正如雨后春笋，蓬勃兴起，使人们可以简捷便利地在线交流、协作、发布、分享、传播信息。社会媒体已成为人们日常生活中不可或缺的一部分，对社会的发展产生了巨大的冲击和影响。如何对海量繁杂的社会媒体进行高效的搜索和分析，是当前学术界和工业界的研究热点。本报告从社会媒体用户与资源建模、社会媒体中对象关系抽取、支持社会媒体的图搜索、社会媒体的群组探测、社会媒体的话题建模、社会媒体上下文及标签信息应用、社会媒体多模态内容分析以及社会媒体搜索的应用等几个方面，较系统地总结了社会媒体搜索与分析技术的发展现状，分析和对比了国内外研究和开发的差距，并对今后的发展趋势提出了预测和建议。

关键词：社会媒体，社交网络，信息检索，数据挖掘

Abstract

With the advent of the Web 2.0, the mushrooming social media, such as QQ, Weibo and WeChat, can facilitate people with more convenient online information communication, collaboration, publishing, sharing and dissemination. The social media has become the indispensable part of people's daily life and has great impact and influence on the development of the society. How to efficiently search and analyze the huge amount of social media data is the major concerns for both academia communities and industry companies. This report focuses on the social media user and resource modeling, entity relation extraction in social media, graph search for social media, community detection in social media, topic modeling in social media, context and tag information in social media, multi-modal content analysis in social media and the application of social media search. Firstly, this report systematically summarizes the state-of-the-art development of social media search and analysis techniques. Then the gap between domestic and foreign research and development is discussed. Finally, this report introduces the predictions and suggestions about the future trends.

Keywords: Social Media, Social Network, Information Retrieval, Data Mining

1 引言

随着 Web 2.0 时代的到来，网络用户已经从被动的“读者”变成了主动的“写者”和“建设者”，Web 2.0 技术的发展使用户可以在线交流、协作、发布、分享和传播信息，于是一种新的技术平台——社会媒体应运而生。社会媒体存在多种模式，各自有其自身的特点，彼此相互独立又具有一定的联系。本部分将介绍社会媒体的定义、分类、作用，以及面向社会媒体的多种搜索模式。

1.1 社会媒体的定义及分类

社会媒体 (social media) 是人们用来分享意见、观点及经验的工具和平台。不同于传统的社会大众媒体，社会媒体作为一种在线交互媒体，让用户享有更多的选择权利和编辑能力以及自行集结成某种阅听社群的功能。因此，具有广泛的用户参与性。

目前，比较流行的社会媒体的形式包括^[1]以下几类：

- 博客 (Blog) 类，一种个人或小组的日记，用以发表和交流个人或小组的心得体会，常用的博客平台有新浪博客、Bloger。有字数限制的简短博客，称为微博 (MicroBlog)，用以分享简短实时信息，如新浪微博、腾讯微博、Twitter 等；
- 维基 (Wiki) 类，网络上众人参与编写的百科全书，常见的百科类平台有 Wikipedia、百度百科等；
- 论坛 (BBS) 类，一种电子公告板，用户在上面可发布信息，分享观点，进行讨论。网络上有大量的论坛，如 55BBS、Discuz！等；
- SNS (社交网络) 类，网络上的人们进行交友的社区，如 QQ、微信、人人网、Facebook、LinkedIn 等；
- 内容社区类，网络上人们进行分享内容的社区，如优酷、豆瓣、YouTub、Flickr 等；
- 播客 (Podcast) 类，个人的网络广播或电视，是一种提供了声音和视频等多媒体形式的博客，如新浪播客、QQ 播客、Podcastalley。

1.2 社会媒体的作用

由于社会媒体使用户拥有了广泛的参与权，用户已经成为信息传播的重要渠道以及人们日常信息传递的主要平台。人们可以利用博客类的社会媒体展示自己的工作、生活、思想、观点等，利用维基类的社会媒体分享自然科学、社会科学及日常生活中的各种知识，利用论坛类的社会媒体对各种社会热点问题和事件进行直接和间接的交流，利用社交网络类的社会媒体进行网上交友并建立各种主题的“圈子”，利用内容社区类的社会

媒体上传，并与他人分享以各种媒体形式表达的信息，如照片、视频、音乐等。此外，微博、微信等一大批新兴社交媒体平台不断涌现，其用户数更是日益增多。

据速途研究院^[2]截至2013年11月25日统计，QQ空间为第一社会化媒体，覆盖比例达89%，并且其用户分享率以13.55%排名第一位，而用户在微信上花费评价时长最长，达40分钟。具体地，主要社交媒体的用户数量、覆盖人群、用户分享率以及用户花费在媒体上的平均时间如图1所示。

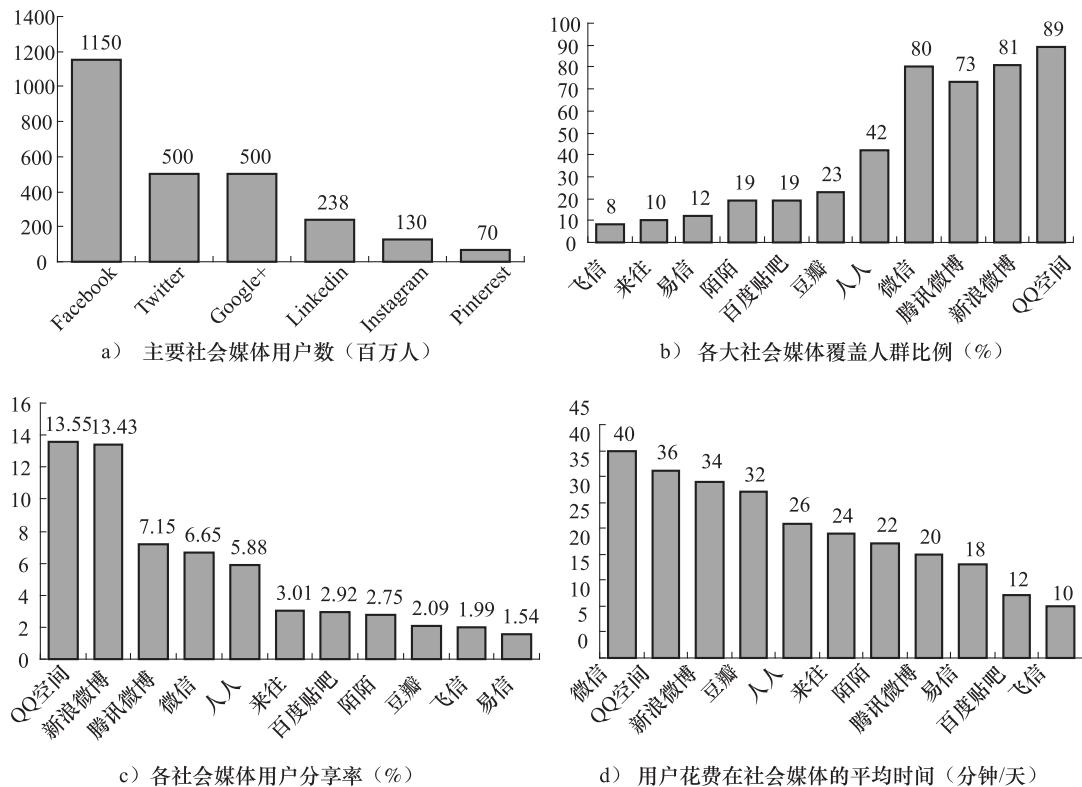


图1 社会媒体统计数据

1.3 社会媒体的特点

与传统媒体相比，社会媒体由于用户的参与度高，呈现出许多新的特点：

- 1) 用户角色的双重性：社会媒体中信息的传播是“众对众”方式，用户不仅是信息的接收者，也是信息的发布者，发布时不仅提供信息，而且在相关网站上建立详细的个人档案并分享这些信息^[3]。
- 2) 用户之间的互动性：一方面，用户之间的相互“关注”、“粉丝”、“回复”等构成直接的互动；另一方面，由于一个用户评价或转发另一个用户上传的信息而形成间接的互动。
- 3) 信息资源的共享性：目前许多主流社交媒体通过注册即可参与，用户产生的内容

大多是公开并可以分享的。

4) 用户参与的社区性: 在社交媒体中, 用户之间由于某个话题形成圈子, 或进一步形成一个交流社区来进行交流。

5) 信息之间的关联性: 一方面, 不同用户发布的信息本身是关于同一地点、事件或话题的; 另一方面, 不同来源的信息在内容、语义、情境等方面构成内在的关联关系。

社会媒体的上述特点, 使社会媒体成为一个数量巨大的用户与资源构成的复杂关系网络。这里的“用户”即社会媒体资源的发布者和使用者(一般需要注册), 包括上传、转载、共享、浏览、关注等一切对社会媒体资源进行过操作的用户; “资源”即用户在社会媒体中发布的原始信息单元及其组合或抽象。原始信息单元如一篇博文、一条微博、一幅图片、一段视频、一首乐曲等, 称之为单一资源; 单一资源中的某个部分称为子资源, 如微博中的图片、视频中的音乐等; 单一资源的有序组合称为复合资源, 如不同用户上传的关于同一景点的多幅图片及若干文本; 通过对单一、复合资源的分析和挖掘, 有望得到综合资源。例如, 对于某个景点, 对用户上传的图片、撰写的博文、发布的微博等进行深入挖掘可构成“旅游综合资源”; 针对某个学术问题, 对相关社群用户的讨论、发表的文章乃至该社群本身的挖掘可构成“学术综合资源”。社会媒体为个性化搜索等网络服务提供了更加丰富的信息资源。因此, 基于社会媒体的搜索技术目前已成为一个新的研究热点。

1.4 社会媒体搜索需求

站在社会媒体搜索服务提供者的角度, 为社会媒体用户提供社会媒体中资源本身、资源组合、资源与资源之间以及用户与资源之间关系的搜索。具体地, 社会媒体搜索系统应提供如下搜索形式:

(1) 搜索与某一资源相关的资源

基于前面关于社会媒体资源的定义, 这种搜索主要基于资源之间的关系来进行。例如, 根据给定的某个文档, 可以搜索与该文档具有相同话题的文档、图像、视频、音乐等; 根据给定的一幅图片, 可以搜索与该图片相似的图片、关于该图片介绍的文档等; 根据给定的一段视频, 可以搜索该视频中的重要片段、其中的故事和人物介绍的文档、视频中的音乐等。进而, 由于资源的不同粒度, 这种搜索还可以包括根据给定的话题搜索相关的资源。显然, 这种搜索包括同类媒体及跨媒体的搜索。

(2) 搜索与某一用户相关的资源

这种搜索是以用户为搜索条件的。例如, 搜索某一用户上传的全部资源, 搜索某一用户转发的信息, 搜索某一用户标注的各种标签。总之, 对于一个给定的社会媒体用户, 搜索该用户所操作的资源。

(3) 搜索与某一资源相关的用户

这种搜索是以资源为搜索条件的。与(1)不同的是, 该搜索是通过资源搜索相关的用户。例如, 搜索转发或评论过某一新闻的所有用户, 搜索上传了某个(些)图片的用

户，搜索为某一视频进行过标注的用户等。与（1）相似的是，这种搜索也包括搜索参与某一话题的用户。

（4）搜索由多种资源构成的综合资源

这种搜索的目标是前面定义的“综合资源”，例如前述的“旅游资源”、“学术资源”等。如前所述，综合资源是通过对单一、复合资源的分析和挖掘而得到的。因此，综合资源的搜索比前面几种搜索更为复杂，一方面表现为搜索的目标不是以简单的形式存在，另一方面表现为搜索需求难以明确表达，这种搜索需要社交媒体资源挖掘、用户搜索意图分析等方面技术的支持。

1.5 与已有搜索模式的区别

目前存在的典型搜索模式可划分为传统的搜索引擎（如 Google、Baidu、Sousou 等）、社交网络搜索（如 Facebook、人人网等）和社会化搜索（如社交网络成员内的搜索）。

（1）与传统搜索引擎的区别

首先，传统的搜索引擎无论搜索的结果是网页、图像、视频、音频还是地图，搜索关键字仍以文本和图片为主。社交媒体搜索应考虑包括上述各种媒体形式的关键字、特别是非文本关键字搜索；第二，传统的搜索引擎搜索时主要针对媒体内容本身，力求准确理解其所表达的语义含义。社交媒体搜索则考虑更广泛的语义信息（如文化、偏好、热点等）。因此，同样的媒体内容，不同用户的操作可能赋予不同的语义信息），更复杂的信息关联（主要是资源与资源、资源与用户关系），形式多样的用户对资源的操作（如信息发布、评论等）；第三，传统的搜索引擎其信息源是 Web 上所有可以公开发布的信息，可以为所有的 Web 用户提供搜索服务。社交媒体搜索由于考虑用户与资源的关系以及用户对资源的操作，同时由于一些社交媒体本身对于信息共享的限制，所以对于每种社会媒体内、外的用户，所提供的搜索服务是有所区别的；第四，传统的搜索引擎提供的是针对媒体内容本身的搜索，为提高搜索效率，主要采用倒排索引等结构。而社交媒体搜索需要对用户与资源、资源与资源关系进行复杂的建模，例如 Facebook 推出的将人（用户）和物（地点、照片、商品等）通过社交关系链接起来的图谱以及 Google、Baidu 等公司面向新兴社交媒体搜索开发的知识图谱。

（2）与社交网络搜索的区别

虽然广义上，社交网络与社交媒体中的内容均为用户生成内容，均具有前述的用户角色的双重性、用户之间的互动性、信息资源的共享性、用户参与的社区性以及信息之间的关联性等特点，但就搜索服务而言，社交网络搜索是以人为中心的搜索，搜索的目的是从社交网络中抽取出人的信息或人与人之间的关系。社交媒体搜索是以内容为中心的搜索，社交媒体搜索虽然考虑了与用户相关的资源搜索，但考虑的是用户所操作（如上传、转发、浏览、分享等）的资源以及操作资源的用户，而非单纯的用户与用户的关系。

（3）与社会化搜索的区别

社会化搜索意为通过搜索形成一个有共同爱好的人际圈子，又通过搜索每个人的爱

好和收藏为用户提供一个更为准确的信息。社会化搜索引擎通常都具备元搜索、收藏、圈子等功能，来满足他们最终达到一个全社会知识共享的概念^[4]。例如：雅虎提供的社会化搜索，是将用户搜索请求，转发给社交网络成员，进而得到答案。社会化搜索是对传统搜索引擎的一种改进。虽然社交媒体搜索在关注用户与资源关系方面与社会化搜索具有一定的相似之处，但社交媒体搜索仍是以内容为中心的搜索，在内容分析方面较社会化搜索更加广泛和深入。

实际上，无论传统的搜索引擎、社交网络搜索、社会化搜索还是社交媒体搜索，虽然在数据源、服务对象、搜索方法和关注点等方面有很大区别，但每种搜索技术均相互借鉴并不断改进，如 Google、Baidu 等传统搜索引擎目前已开发了具有社交媒体搜索的功能。

2 国际研究现状

社交媒体已成为一个巨大的用户与资源构成的复杂关系网络，其内容包括多种媒体（模态）形式。为有效地进行社交媒体资源、用户及其关系的搜索，首先应该构建支持社交媒体搜索的模型。构建模型则需要在大量无结构的社会媒体内容中抽取不同层次的对象，并分析对象之间的关系。进而，根据社交媒体的特点，群组关系的探测以及话题建模是对社交媒体资源与用户模型的进一步扩展。表征社交媒体资源及用户关系模型需要复杂的数据结构，其中以图结构为主。因此，基于社交媒体模型的搜索实际上是以图搜索技术为支撑的，而群组探测和话题建模的结果为社交媒体搜索提供了更多的搜索条件和更丰富的搜索结果。同时图中的节点包括用户和多模态资源，因此社交媒体建模和搜索还需要多媒体内容分析。此外，社交媒体中的媒体上下文线索及大量的社会标签是实现搜索的一个可利用的资源。上述技术构成了社交媒体搜索的主要相关技术，其关系如图 2 所示。

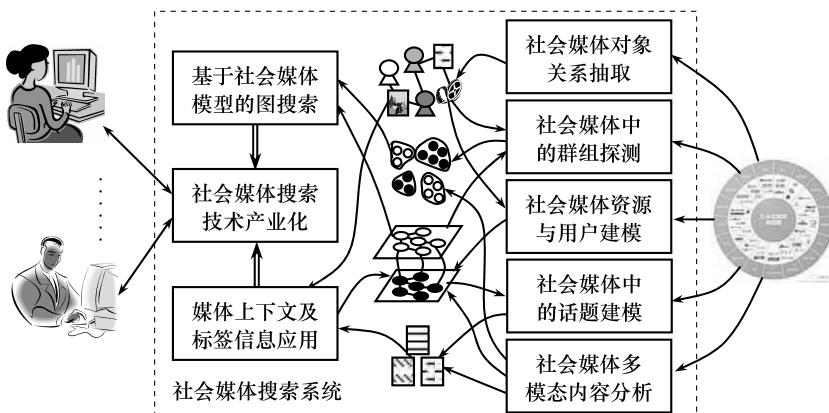


图 2 社会媒体搜索相关技术

如图 2 所示，本节将从社交媒体用户与资源建模、社交媒体中对象关系抽取、支持社交媒体的图搜索、社交媒体中的群组探测、社交媒体中的话题建模、社交媒体上下文及标签信息应用、社交媒体多模态内容分析以及社交媒体搜索的应用——系统、产品和专利几方面阐述研究现状。

2.1 社会媒体用户与资源建模

如前所述，社交媒体已成为一个巨大的用户与资源构成的复杂关系网络，对其建立合理的模型，并进行相关的用户与资源挖掘，是社交媒体搜索的前提。基于此，相关学者在社交媒体用户与资源建模方面进行了研究，建模的对象包括各种社交媒体网络的用户和资源，采用的数据结构以复杂图结构为主。

美国南加利福尼亚大学与香港中文大学^[5]提出特征、推特、用户三重图模型，建模三者之间的互依赖关系，通过三重图聚类，研究用户层和推特层的情感分析，获得了更佳的动态情感分析质量；美国波音研究与技术公司^[6]应用张量空间模型对推特数据进行建模，描述数据中诸如“关注”、“好友”、“回复”、“引用”、“评论”等社会关系及话题的演化过程；澳大利亚国立大学^[7]提出一种基于概率方法的网络传播模型，将社交媒体网络中异构和结构化的关系相结合，研究新闻、社会网络、博客媒体之间的新闻传播，发现了新闻在不同的媒体类型中的不同影响；美国南加州大学洛杉矶分校^[8]提出一个三重图的产生式模型，描述用户、资源和标签之间的关系，发现能够描述这种关系的显著话题，基于此为新的用户推荐合适的标签和相关的用户；美国谷歌公司、雅虎研究院及卡内基梅隆大学^[9]提出一个集成了位置信息和消息内容的产生式模型，将分布的位置、话题和用户特点相结合，建模社交媒体中用户的地理信息，从而为推特提供话题模型，获得特定话题的位置，推断话题位置的潜在分布，提供位置与话题相结合的层次模型，并推断用户的个性化偏好。西班牙马德里卡洛斯三世大学^[10]提出一种数据驱动的模型，基于连续时间 Hawkes 过程，利用微博网站的用户活动数据预测商品的竞争动态；荷兰代尔夫特理工大学^[11]使用超图及超图中的线图建模具有重叠社群的社会网络，用超图的节点表示社群，而线图的节点为一个个体（线图节点），超图的超边表示可以加入多个社群的线图节点。因此，社群由这些加入的线图节点及它们之间的链接构成，并成为超图中的一个节点。

此外，针对社交媒体的建模，还包括犯罪模型的构建^[12]、社会面貌模型的构建^[13]、用户行为概率模型构建^[14]，特别是社会影响力传播模型^[7, 15, 16]等。

在社交媒体资源与用户建模的研究中，其模型结构涉及超图、多层次图、张量空间模型等，这些均非简单的数据结构。因为对社交媒体的资源与用户关系建模，将带来如下问题：首先，表现这些资源的数据本身大多是非结构化的；其次，不同资源数据之间是内容异构的；第三，这些非结构化数据的特征是高维的；第四，用户对资源的操作是多样的。

若准确而合理地表达不同资源之间的关系，并有效地支持各种应用，已非简单的数

据结构所能胜任。这种情况下，能够表达结构和语义信息的复杂图模型是一种有效的模型结构。

2.2 社会媒体中对象及关系抽取

社会媒体资源和用户建模是社会媒体搜索的前提。而社会媒体模型的构建，大多基于复杂的图模型。图模型包含顶点和边，通常将用户及各种资源或其特征建模成顶点，而将它们的关系建模成边。因此，基于复杂图的社会媒体建模中，首先是对象的识别和抽取，然后是这些对象之间关系的度量。基于此，相关学者早在社会媒体中对象抽取机器关系分析方面进行了研究，抽取的对象包括同构和异构的，其间关系的度量也基于不同的方法和准则。

希腊亚里士多德大学^[17]将多人感兴趣的事件视为重要事件，集成了命名实体识别、动态话题映射发现、话题聚类以及峰值检测技术，从具有时间戳的 Web 文档流中检测重要事件；瑞士洛桑联邦理工学院^[18]基于用户在社会网络中发布的内容来构建 user profile，从而利用用户在社会网络中的活动历史抽取用户声称内容中的语义特征，基于语法和语义进行实体消歧；美国 Walmart 实验室、LinkedIn、Wisconsin-Madison 大学，Cambrian Ventures 公司^[19]使用基于维基百科的实时知识库，在社会媒体数据中进行实体抽取、链接、分类和标注。

无论前述的对象、事件，还是本文中所称的资源和用户，它们构成了社会媒体图模型中的顶点，而对象之间的相似性度量和相关性分析是构建顶点之间关系（即边）的重要基础。为此，新加坡南洋理工大学^[20]提出一种在线多模态深度相似性学习框架，针对每种单一模态学习一种非线性转换函数，在此基础上去学习发现多种模态的最优组合，与通常“模态”的定义不同，这里所述的模态是图像的不同类型的特征；雅典国立技术大学^[21]研究社会事件检测问题，为实现相似多媒体项的聚类，以中餐馆顾客为例，提出以题目、描述、标签、空间位置（经纬度）、签到时间等属性构成的顾客实体之间的相似性度量方法；美国斯坦福大学和康奈尔大学^[22]则是从另一个角度研究用户之间的相似性，以 Wikipedia、Stack Overflow 和 Epinions 作为研究网站，提出根据用户对网站内容的贡献程度和用户与其他用户的交互情况度量用户之间的相似性。

此外，由于社会媒体资源是多媒体（模态）的，因此，诸如图像^[23]、音乐^[24]、视频^[25,26]等媒体对象之间的相似性度量均可作为社会媒体对象之间关系度量的基础。

由于社会媒体中包括用户和多模态资源，因此上述工作中，抽取的对象包括不同类型的用户和不同模态的资源，而对象之间的关系度量同样涉及相同模态对象或实体之间的相似性度量，更涉及跨模态对象之间的相关性分析。这里的“模态”可以视为不同的媒体，也可以视为相同媒体的不同特征，上述工作也涵盖了这几种模态之间的相似性度量问题。在不同媒体资源、用户及用户对资源不同操作构成的社会媒体复杂图模型中，跨模态对象之间的相关性分析和相似性度量尤为重要。

2.3 支持社交媒体搜索的图搜索

将社交媒体资源建模成复杂的异构网络图，因此，社交媒体搜索实际上就是这种异构网络图上的搜索。单一资源的搜索为顶点搜索，而用户与资源、资源与资源关系的搜索就是图搜索。因此，图搜索成为社交媒体搜索的理论基础，相关学者在此进行了研究，内容包括子图、对象（点）、对象间关系（边）等的搜索。同时，由于对象包括同构和异构的，因此图的搜索也包括同构、异构图上的搜索。

在图搜索方面，Facebook 公司^[27]提出一个在线的、基于内存的、社会图感知索引系统，以支持具有上千种商品实体和数百亿用户之间数万亿条边的图搜索服务，通过 Apply 和 Extract 操作实现用户对商品的搜索；法国 Cergy-Pontoise 大学^[28]在推荐系统的开发中，认为用户之间通过协同社会网络相互关联，而用户与项之间通过语义类别关联，进而改进了图上深度优先搜索策略，实现基于节点 - 边的推荐和基于节点的推荐；巴西蒙特卡洛法大学、卡塔尔计算研究院和美国伦塞勒理工学院^[29]提出一种精炼的在线搜索方法，处理在大图中的可达性查询问题，回答“给定一个图，在两个任意定点之间是否存在一条路径”的查询；美国加利福尼亚大学欧文分校、爱尔兰 IBM 研究院^[30]提出加权的最佳优先搜索策略，将传统的最佳优先搜索方法扩展到图上的搜索。

社交媒体建模是构建表达用户、资源及其关系的模型，由此构建的模型其节点类型是异构的，许多学者将这类图成为“异构信息网”^[31]。在异构信息网络搜索中，将伴随着信息融合的问题。在该领域，美国伊利诺伊大学香槟分校^[32]在多类型实体和链接构成的异构信息网络搜索中，提出采用基于元路径的排序模型，实现用户指导的用户查询，解决缺少上下文而导致的查询语义含糊的问题；他们还提出了异构信息网络中基于元路径的 top-k 相似性查询方法^[33]；澳大利亚墨尔本大学和维多利亚研究院^[34]结合各种社交媒体网络中的搜索特点，研究异构信息网络中的搜索问题，提出一种改进的基于元路径的相似性度量方法，将相似性与时间信息相结合，从而改进了搜索质量。

上述工作中，无论建模时的信息融合、还是搜索或推荐时在线的信息融合，均涉及图上的游走问题，而前述的搜索过程更是在复杂的大图中寻找子图、路径或节点。因此，图搜索的效率是提高社交媒体搜索效率的关键。

2.4 社交媒体中的群组探测

群组探测旨在将社交媒体划分为一个个群组，并且组内关系密集而组间关系稀疏。将这种定性的认识进行量化是基于度量优化的群组探测算法的主要思路。目前，用于评价群组划分结果好坏的度量主要包括归一化割（Normalized Cut, Ncut）^[35]、模块度（Modularity）^[36,37]、最小描述距离（Minimum Description Length, MDL）^[38~40]和互信息（Mutual Information）^[41]等。基于模块度优化的群组探测算法就是将这些度量作为目标函数，通过优化这些度量探测出社交媒体中的群组。然而归一化割

和模块度最优化问题都是 NP 难问题，目前还不存在多项式时间复杂度的精确算法，已有的文献里提出了许多近似或启发式算法。谱聚类算法^[42]是重要的群组探测近似算法之一，该算法将群组探测这一离散最优化问题进行松弛化，转变成拉普拉斯矩阵的特征向量求解问题，最后在特征向量上执行 k -means 聚类算法^[43]得到近似最优解。

概率生成模型（Probabilistic Generative Model）通过定义网络中链接的生成过程并利用概率统计模型对节点所属的群组进行联合推理。典型的基于概率生成模型的群组探测算法包括随机块模型（Stochastic Block Model, SBM）^[44]和混合隶属度随机块模型（Mixed Membership Stochastic Block Model, MMSB）^[45]。SBM 模型首先对网络中的每个节点，根据多项式分布对该节点所属的群组进行抽样。然后，根据每对节点在第一步中抽样出的群组对和伯努利分布，对该节点对之间的链接关系进行抽样。MMSB 在 SBM 的基础上建立软聚类模型。

现实中社会媒体是随时间动态变化的，为了更好地理解和挖掘动态社会媒体中群组随时间的演化过程，目前动态群组探测也受到广泛的研究。在考虑时间方面，常用的方法是将时间按照预先定义的时间间隔划分成时间片，每个时间片对应社会媒体的一个快照。雅虎公司^[46]提出演化聚类（Evolutionary Clustering）方法，在保证每个快照的群组探测结果质量的同时尽可能保证相邻时间片的群组探测结果的一致性。基于演化聚类的思想，美国的 NEC Laboratories^[47]提出演化谱聚类算法，亚利桑那州立大学、美国的 NEC Laboratories 以及雅虎公司^[48]运用隐马尔可夫模型和矩阵分解技术分析社会媒体中的群组及其演化，伊利诺伊大学香槟分校^[49]提出基于质点和密度的演化聚类方法。密歇根州立大学和美国的 NEC Laboratories^[50]在随机块模型的基础上提出动态随机块模型（Dynamic Stochastic Block Model, DSBM）。卡内基梅隆大学和 IBM 公司^[39]提出 GraphScope 算法对动态二部网络进行挖掘，该算法可以探测出社会媒体中群组结构发生明显变化的时间点并且挖掘每个时间段内的群组。伊利诺大学芝加哥分校和中山大学^[51]主要是在有权的图流里面进行动态群组探测。明尼苏达大学、IBM 公司以及伊利诺大学芝加哥分校^[52]利用先验知识进行群组探测。

Statistical and Biological Physics Research Group of the HAS^[53] 利用派系渗透方法（Clique Percolation Method, CPM）研究动态社会媒体中群组的演化，类似的，Ohio State University^[54]首先将每个时间片的群组抽取出来，然后通过比较相邻时间片的任意一对群组的大小以及它们之间的重叠情况来发现群组演化事件。都柏林大学^[55]将动态群组（Dynamic Community）定义为按时间先后顺序排列的群组的集合，利用类似 [53] 和 [54] 的方法定义群组演化事件，并提出动态群组匹配、发现和排序算法。伊利诺大学芝加哥分校和南加州大学^[56]认为群组在一段时间内且相对稳定，将群组与快照中的临时组区别对待，然后将动态社会媒体中的群组探测问题转化为最小着色问题。亚利桑那州立大学、美国伊利诺伊大学香槟分校以及清华大学^[57~59]则对异构网络中的动态群组探测问题进行了研究，异构网络定义为含有多种类型的对象及链接关系的网络。伊利诺伊香槟分校和美国的 NEC Laboratories^[60]提出了增量式谱聚类算法，然而该算法不能探测重叠群组（Overlapping Communities）且存在累积误差（Accumulating Error）。

2.5 社会媒体中的话题建模

作为重要的社交媒体挖掘技术，话题建模在过去的十年中受到了广泛的关注^[61,83]。最早的话题建模技术是概率潜语义分析（Probabilistic Latent Semantic Analysis, PLSA）^[61]。PLSA 是一种概率生成模型，它通过在文本和单词之间引入话题层对文本的生成过程进行建模。具体来说，对每个文本里的每一个位置，首先根据该文本的话题分布对该位置的话题进行抽样，然后根据话题的单词分布对该位置的单词进行抽样。然而，PLSA 存在过拟合（Over Fitting）问题，针对该问题加利福尼亚大学^[62]提出潜狄利克雷分配（Latent Dirichlet Allocation, LDA）模型，该模型假设文本的话题分布服从参数为 α 的狄利克雷先验分布，所有文本共享同一个先验话题分布，减少不同文本之间的话题分布的差异，能够降低文本数据中噪声对模型的影响，从而提高模型的泛化性能。如果说 PLSA 是话题建模的起源，那么 LDA 的提出是话题建模技术被广泛研究的起点。话题建模的好坏可以从泛化性能和实际应用效果两方面进行评价。Peplexity^[63]被广泛用于度量话题建模的泛化性能，Peplexity 越小，模型的泛化性能越好。话题建模还可以应用于文本聚类、分类、信息检索^[64]和推荐。话题模型在这些应用上的表现越好，表明话题模型本身也就越好。经过十几年的发展，话题模型已经有许多的变种，主要包括结合链接关系的话题模型、动态话题模型、作者话题模型、有监督的话题模型、将群组和话题相结合的模型等。

由于基本话题模型只考虑了文本信息而忽略了社会网络中的链接关系，其泛化性能受到了限制。文本信息中通常包含噪声信息，而链接关系可以帮助模型过滤掉一些噪声。结合链接关系的话题模型旨在将链接信息结合到话题模型中，从而提高话题模型的泛化性能和实际应用效果。伊利诺伊大学香槟分校^[65]提出一种基于网络正则化的话题建模（Topic Modeling with Network Regularization, TMN）框架，该框架将文本的话题分布按照网络结构进行正则化处理，假设网络上相邻的文本具有相似的话题分布。在 TMN 框架下，提出了 NetPLSA 算法，并通过群组探测的效果及话题的凝聚性（Coherence）表明 NetPLSA 比没有考虑链接的话题模型的性能更优越。伊利诺伊大学^[66]提出 iTopic 模型，运用马尔可夫随机场（Markov Random Field, MRF）对相邻文本之间的话题分布依赖关系进行建模。伊利诺伊大学提出的热点事件跟踪（Popular Event Tracking, PET）^[67]假设用户的兴趣受到其邻居用户的兴趣的影响，这一点在基本思想上和 TMN 是类似的。伊利诺伊大学^[68]将话题模型扩展到异构信息网络（Heterogeneous Information Networks）中。普林斯顿大学^[69]提出关系话题模型（Relational Topic Model, RTM），在链接和话题之间建立依赖关系。其基本思想是利用链接关系来指导话题建模过程，将每两个节点之间的链接关系建模为这两个节点的话题分布的对应元素乘积（element-wise product）的逻辑回归函数。斯坦福大学^[70]提出 TopicFlow 模型，模型利用网络流模型对文本之间的相互影响进行建模，并将网络流模型和话题模型进行有机的结合，对含有超链接结构的文档网络进行统一建模。

除了在话题模型中加入链接关系，在话题模型中考虑的另一个重要的信息是时间特征。语料库中的文本往往都与一个时间相关联，如博文的发布时间、论文的发表年份等。在话题建模过程中考虑时间信息，能够跟踪话题的演化过程。普林斯顿大学^[71]提出动态话题模型（Dynamic Topic Model, DTM），将时间划分成一个个离散的时间片，利用多项式分布的自然参数来表示话题，将话题在相邻时间片的依赖关系建模为高斯分布。普林斯顿大学^[72]研究了连续时间的动态话题模型（continuous-time Dynamic Topic Model, cDTM），不需要将时间进行离散化，其基本思想是利用布朗运动（Brownian Motion）对贯穿文本序列中的隐含话题进行建模。在建立相邻时间片的依赖关系时，DTM 和 cDTM 都是基于马尔可夫假设，即当前话题只依赖于前一个时间片的话题，并利用状态空间模型进行建模，杜克大学^[73]则利用线性模型进行建模，将当前时间片的话题分布建模为前一个时间片话题分布的线性函数。卡内基 - 梅隆大学^[74]则基于马尔可夫假设解决话题随时间的出现和消失的问题。NTT 通信科学实验室提出的多尺度动态话题模型（Multi-scale Dynamic Topic Model, MDTM）^[75]考虑到话题产生影响的周期有长有短，在动态话题模型的基础上考虑话题跨多个时间尺度（Multiple Time Scale）的依赖关系。马萨诸塞大学^[76]提出 TOT（Topic Over Time）对文本和文本的生成时间同时进行建模。利用传统话题模型类似的思想，TOT 用多项式分布对话题的单词分布进行建模，由于时间是连续的，TOT 选用贝塔（Beta）分布对话题的时间分布进行建模。芝加哥大学^[77]利用在线非负矩阵分解（Online Nonnegative Matrix Factorization）方法研究非结构化文本流中的话题出现和演化过程。

话题模型本身是无监督的机器学习算法，然而目前话题建模也被广泛应用于一些预测任务，在训练过程中利用预测任务的响应变量指导话题模型的结果。总的来说，其基本思想是在预测任务的响应变量（Response Variable）和话题模型的隐含变量之间建立函数依赖关系，然后利用训练集学习出这些依赖关系的参数，最后用学习出的参数进行预测。普林斯顿大学^[78]提出有监督的话题模型（Supervised Topic Models），在预测变量和话题之间建立概率依赖关系，并利用该模型预测用户评论文本中的情感倾向和 Web 页面的流行度（Web Page Popularity）。普林斯顿大学^[79]又将协同过滤（Collaborative Filtering）和话题建模相结合进行科技论文推荐，即预测用户可能感兴趣的论文，其基本思想是在每个文档的话题分布（Per-document Topic Distribution）和文档的潜在向量（Document Latent Vector）之间建立依赖关系。理海大学^[80]在动态话题模型中加入词频（Term Volume）信息实现对词频的预测，其主要思想是将词频建模为话题的线性函数。卡内基梅隆大学^[81]提出时变用户模型（Time Varying User Model, TVUM）对用户的动态兴趣进行建模，并利用用户兴趣对用户的广告点击行为进行预测。卡内基 - 梅隆大学^[82]又利用有监督的话题模型预测多社区（Multi-community）用户的情感极性（Sentiment Polarity），即正面或负面情感，以及评论数目（Comment Volume）。普林斯顿大学^[83]利用有监督的话题模型预测立法机构或立法者对法案的表决（Voting）行为。

2.6 社会媒体上下文及标签信息的应用

社交媒体平台具有广泛的用户参与性，因此，大量的社会标签成为社交媒体资源的一大特色，而众包^[84]、分众分类^[85]等技术的出现使标签更加丰富。在应用社交媒体的上下文线索及标签信息进行搜索方面，相关学者进行了研究，内容涉及利用标签建模和搜索、以不同媒体形式的“共现”关系作为线索以及从媒体资源中抽取上下文关系等。

在基于标签的社会媒体搜索中，加拿大渥太华大学^[86,87]利用社会标签作为一种偏好指示器，构建两个模型：潜在的标签偏好模型和潜在标签标注模型，研究用户如何为资源赋予标签的问题，提出一种新的基于标签的个性化搜索模型；荷兰代尔夫特大学^[88]在社会媒体搜索中，通过随机游走进行排序，其中考虑协同标签对用户兴趣的影响，从而更好地发现与用户兴趣相关的内容；芬兰于韦斯屈莱大学^[89]分析社会标签中的语义信息，基于音乐社会媒体中的标签计算音乐情感。

此外，同一用户操作的资源、同一资源中不同的媒体内容又构成了资源之间的媒体上下文线索（可视为“共现”），这些线索进一步延伸，又将获得更多的线索。这种媒体上下文线索不仅是基于情境的搜索的主要依据，而且是不同模态资源的相似性度量的一个重要依据，为社会媒体的建模和搜索提供支撑。

在媒体上下文线索的应用方面，美国微软研究院、佐治亚技术学院、丹麦奥尔堡大学^[90]提出一个讨论图模型，以捕获社会媒体联系中的结构特征以及相关于讨论的上下文，如讨论的参加者、时间、地点和内容等，其中重点考虑“共现”信息分析，抽取诸如“位置↔活动”、“药品↔副作用”等特定领域的共现关系用以建模；美国雷克塞尔大学^[91]应用共现分析方法识别与种子术语频繁共现的术语，从而自动地在消费者生成内容中识别消费者的健康状况；英国谢菲尔德大学和丹麦奥尔胡斯大学^[92]结合时间上下文、空间上下文以及时空上下文，研究社会媒体中以数据为中心的上下文感知的搜索和分析，提供新鲜的和更加相关的搜索结果。

2.7 社会媒体中的内容分析

内容分析是基于内容搜索（CBIR）的基础，在传统的搜索系统、特别是多媒体搜索中起着重要作用。虽然社会媒体中的标签和上下文信息对其搜索提供了重要的帮助，但传统的多媒体内容分析在社会媒体检索中仍具有其不可替代的作用。因此，许多相关学者将多媒体内容分析技术应用于社会媒体的搜索。

在多模态内容分析方面，新加坡国立大学、中科院智能机械研究所和新加坡国立大学^[93]基于用户输入的图像关键字分析用户搜索意图，通过对图像不同层次的特征分析及反馈机制，返回给用户所需的图像；罗马尼亚克拉瓦约大学^[94]提出一个在格结构上构建的可视超图的最大生成树，进行图像的“色彩 - 色度 - 亮度”的彩色空间表达，以实现基于超图结构的图像划分及边界抽取；美国俄亥俄州立大学^[95]将相同分割标签的若干像

素视为超像素，超像素构成超图的顶点，通过连接超像素及其近邻构成超边，每个超边以一个概率与其所包含的那些顶点相关，基于构成超像素的图像特征确定超边的权重，以这样的机制建模图像并实现图像分割；新加坡国立大学与中科院重庆绿色智能技术研究院^[96]获取社会媒体中各种女性面部和发式图像，提取其中化妆品特征，构建一个化妆推荐系统，提出一个多树结构的超图模型来探索各种化妆面部图像的高层美丽属性、中层相关于美丽的属性以及低层图像特征，基于用户输入的短发、素颜正脸图像，为其推荐最合适发式和化妆模式，并展示合成结果。

此外，已有较长研究历史的基于内容的图像^[97,98]、视频^[99,100]等各种媒体（模态）的检索技术仍在不断发展，这些技术同样可以应用于多媒体（模态）社会媒体的检索。

在社会媒体中，由于前述媒体上下文线索和社会标签为社会媒体搜索提供了便捷的渠道，很多时候，似乎不需要内容分析也能取得不错的搜索结果。甚至在 ACM Multimedia 2012 国际会议上还引发了在社会媒体环境下“内容已死”与“内容万岁”的讨论^[101]。但社会媒体搜索总体上是以资源为中心的，社会媒体搜索包含了非文本关键字搜索的需求，而且同一模态资源之间的相似性度量更需要这种精细的内容分析。因此，内容分析仍然是社会媒体搜索不可或缺的技术。

2.8 社会媒体搜索技术的产业化

在社会媒体搜索技术研究的基础上，随着社会媒体的普及，越来越多的科研院所和企业公司设计开发了新型的面向社会媒体的搜索系统，其中一部分已经申请专利，甚至投入商业使用。

2013 年 1 月，美国 Facebook 公司推出一种全新的社会媒体搜索工具 Facebook Graph Search (FGS)^[102]。在传统的 Web 搜索系统（例如 Google）中，用户搜索的是网页，网页排序顺序取决于该网页在整个互联网上被超链接引用的次数。而 FGS 将人（用户）和物（地点、照片、商品等）通过社交关系链接起来，形成“图谱”。在 FGS 中，用户不仅可以搜索传统的互联网内容（新闻、文字、地图等），还可以检索传统搜索引擎无法索引的，与社交密切相关的內容。例如，使用 Google 可以回答这样的检索“在北京的西餐馆”，而 FGS 中可以提交这样的检索“在北京被我的女性朋友推荐过的西餐馆”。可以看出，FGS 强调用户本人在相关社会媒体中的相关检索结果，相关技术已经申请专利^[103]。而这些是传统 Web 搜索引擎无法实现的。

2012 年 5 月，Google 公司发布了知识图谱（Knowledge Graph）^[104]。与 FGS 强调人、物以及他们之间的社交关系不同，Google 公司的知识图谱是由主题实体以及它们之间的联系组成，目的是为用户提供有完整知识体系的搜索结果。例如，当搜索“玛丽·居里”时，用户不仅可以获得这个关键词的所有相关内容，还能获得“居里夫人”的详细生平介绍。实际上 Google 已经将“玛丽·居里”、“居里夫人”以及相关网页形成图谱结构，并采用自然语言理解技术对检索内容进行语义理解，进而提高搜索结果质量。

作为最知名的社会媒体之一，Twitter 具有非常强大的搜索功能^[105]。除了以关键词

的方式检索热点话题之外，Twitter 还支持对用户和指定地点的搜索。同时，在搜索过程中，Twitter 会根据用户的搜索关键词，以及用户已有的好友、标签等信息，为用户推荐新的朋友。

在相关专利申请方面，Flashback 公司申请了多社交媒体源搜索结果的集成展示的专利^[106]。IBM 公司开发了相应技术，可以在社交媒体中基于文字和图片挖掘发生的事件，对事件的图片建立相应的语义概念，并基于这些语义概念进行基于内容的检索与展示，相关技术已经申请专利^[107]。

3 国内研究进展

在社交媒体搜索技术的研究中，无论国外还是国内均涉及图 2 所示的主要相关技术。因此，本节同样基于这些相关技术综述国内研究进展情况。

3.1 社会媒体用户与资源建模

与国外相关技术研究相似，国内学者在社交媒体用户与资源建模方面主要研究了社交媒体中对象及其关系的建模方法、图像和上下文关系建模、新闻与评论关系建模等问题。

香港城市大学、理光软件研究中心、微软亚洲研究院^[108]提出将社交媒体对象等价地视为“实体”，构造一个多层次图，将这些异构实体组织成多个层次，并对同一层次的边（层内边）和不同层次图的边（层间边）分别采用不同的加权方式，从用户的社会行为（如标注、评论以及加入社群等）中提取相关信息为边的权重赋值；浙江大学^[109]采用超图建模社交媒体中图像与其他上下文之间的关系，为这种超图定义了同构超边和异构超边，将传统的基于图的谱映射方法扩展到超图中以加速社会图像的相似性搜索；清华大学、合肥科技大学、新加坡国立大学、新加坡管理大学、悉尼科技大学^[110]根据图像的标注和视觉内容中分别生成词袋和视觉特征袋，基于此构建一个超图以建模图像之间的联系；西南财经大学^[111]采用图模型对评论间的关系以及评论与原始新闻间的关系进行建模，捕捉用户关注点的动态变化，抽取话题模式。

3.2 社会媒体中对象及关系抽取

在社交媒体对象及关系抽取方面，国内学者主要研究了社交媒体中的实体搜索、未知语义实体抽取、多样性语义度量、跨媒体特征相似性度量等问题。

香港城市大学、理光软件研究中心、微软亚洲研究院^[108]将用户和资源均视为实体，研究实体间关系，并基于层次图模型研究实体搜索；华东师范大学、西北大学^[112]通过集成统计特征、决策树、支持向量机算法，应用这种集成的分类方法抽取文本中未知的

语义实体；清华大学、首都医科大学、新加坡管理大学和美国伊利诺伊大学香槟分校^[113]研究社会媒体用户所构成的社会网络中一个节点与对等节点连接方式的多样性问题，提出捕获多样性语义的度量标准，在社会媒体网站中获得各种类型的朋友、合作等关系；北京大学^[114]提出一种支持跨媒体信息检索的异构媒体对象的相似性度量方法，该方法探索一种结合了原始的低层特征空间和第三公共空间特点的 tri 空间，基于该空间进行不同媒体对象的相似性度量；北京大学与新加坡国立大学^[115]提出社会媒体搜索中融合多特征及其相关性的方法进行相似性度量，其特征涉及正文特征、视觉内容特征和用户特征，分别以这些特征为节点构建特征交互图，并考虑同模态节点边和跨模态节点边的定义和构建；广西师范大学与澳大利亚昆士兰大学^[116]提出一种跨模态哈希方法，在对每种模态的数据进行聚类基础上，将得到的数据表达转换成普通二进制子空间，使所有模态的二进制编码是“一致的”和可比较的，同时输出针对所有模态的哈希函数，用于将未知数据转换成二进制代码。

3.3 支持社会媒体搜索的图搜索

在支持社会媒体搜索的图搜索研究方面，国内学者主要研究了社会媒体中社会关系子图搜索、社会关系子树查询、用户兴趣搜索、用户偏好搜索等问题。

清华大学^[117]研究在基于人与人之间关系构建的社会网络中的子图搜索问题，实现“返现包括输入的 k 个人的子图”的搜索，满足用户“寻找求职或求学的推荐人”、“寻找某方面专家即联系方式”等方面的搜索需求。东北大学^[118]应用多模态异构数据构建社会关系图，基于社会关系图通过子树的查询和回答实现图像搜索；台湾交通大学和台湾政治大学^[119]的异构信息网络的节点包含文章、作者、术语等类型、反应多种链接关系，通过执行具有重启功能的随机游走过程，搜索文章和其他类型的实体对象；浙江大学^[120]提出基于潜在狄氏分布（LDA）的贝叶斯层次方法，通过在公共的话题空间建模不同领域的文档和用户兴趣，并学习面向该文档和用户兴趣的话题分布，融合媒体描述、用户生成内容文本和点击率等信息实施跨领域的推荐；香港浸会大学、电子科技大学和淘宝软件公司^[121]在社会媒体推荐中，基于矩阵分解的方法，对推荐的项、组合朋友进行融合，在推荐项目时融合朋友关系和组成员，推荐组时融合朋友关系和用户 - 项偏好，推荐朋友时，融合组成员及用户 - 项偏好。

3.4 社会媒体中的群组探测

在群组探测方面，相较于国际的研究，国内的有关研究成果要相对较弱。传统的基于度量优化的群组探测，主要有模块度（Modularity）^[122~124]和谱聚类算法。基于模块度优化的群组探测算法就是将这些度量作为目标函数，通过优化这些度量探测出社会媒体中的群组。谱聚类算法^[125]是重要的群组探测近似算法之一，该算法将群组探测这一离散最优化问题进行松弛化，转变成拉普拉斯矩阵的特征向量求解问题。吉林大学^[126]提

出了基于节点距离相似度的社区挖掘算法。华中科技大学和海军航空工程学院^[127]介绍了一种基于共同好友数和节点邻居信息的社区结构发现算法；并以节点 Q 值为衡量标准，判断是否将该节点加入到初始社区中，最后根据节点邻居所在初始社区信息确定最终的社区划分。福建师范大学和广西师范大学^[128]提出了一种基于多目标粒子群优化的网络社区发现算法 MOCD-PSO，它选取模块度 Q、最小最大割 MinMaxCut 与轮廓（silhouette）这 3 个指标进行综合寻优。

针对社交媒体的动态变化性，华中科技大学^[129]研究了群组变化点检测算法，检测群组发生显著变化的时间点并对同一时间段内的群组实现快速更新。针对社会网络快速更新的特点，华中科技大学^[130]研究增量式 K- 派系（K-Clique）聚类问题。当社会网络发生变化时，尽可能缩小群组更新的范围，实现群组的局部快速准确的更新。西北工业大学^[131]通过对信息流动的分析来发现联系紧密、兴趣相近的节点集合，以实现动态的社区发现。燕山大学^[132]提出一种给定阈值的 α 关系社区概念。国防科学技术大学^[133]提出了一种可并行分解的层次化动态社区发现算法。

北京交通大学^[134]基于 GSB 模型设计一种快速算法，更快地发现网络的广义社区。扬州大学和南京大学^[135]提出了一个基于蚁群优化的二分网络社区挖掘算法。中国科学院^[136]在信息论的框架下，提出了一种基于信息瓶颈的社区发现方法。国防科学技术大学^[137]提出了一种基于带权图并行分解的层次化社区发现方法，该方法采用图划分的方式定义社区结构，并在这种社区结构之上实现了社会网络社区发现并行算法。北京大学^[138]结合用户的兴趣和社会关系进行群组探测，首先利用用户使用的资源的相似度来反映用户的兴趣相似度，然后利用聚类算法根据用户的兴趣进行聚类，最后利用社会关系进行群组扩展。清华大学^[139]将群组分为核群组（Kernel Community）和附属群组（Auxiliary Community），核群组由某个领域有影响的用户组成，而附属群组由依附在核群组周围的用户组成。

3.5 社会媒体中的话题建模

在话题建模方面，国内的研究也有不俗的成果。国内的成果大部分集中于作者话题模型、结合链接关系的话题建模、考虑稀疏性的模型、新闻话题的挖掘方法、话题检测与发现和话题跟踪等方面。

在结合链接关系的话题建模方面，华中科技大学^[140]提出了基于排序的话题模型（Ranking based Topic Model, RankTopic），该模型改变传统话题建模将文本看做同等重要的劣势，通过引入文本的链接重要性来改进话题建模性能。不仅如此，华中科技大学^[141]提出的无穷群组话题模型（MEI）通过引入狄利克雷过程混合（Dirichlet Process Mixture, DPM）模型和分层狄利克雷过程（Hierarchical Dirichlet Process, HDP）达到自动探测群组和话题个数的目的。北京大学、高可信软件技术教育部重点实验室^[142]考虑了用户发表内容、用户之间的关系信息，利用话题传播、社区形成和用户影响力之间的关联性，提出了一个基于 LDA（Latent Dirichlet Allocation）的集成话题发现、社区发现和

用户影响力分析的统一模型 ACT-LDA (author-community-topic LDA)。该模型采用变分推理的方法解决推理问题。

清华大学、IBM 中国研究院^[143]在作者话题模型的基础上，考虑论文的会议信息，提出作者会议话题模型 (Author Conference Topic Model, ACT)，同时对论文的作者和会议进行建模，并将该模型应用于论文、会议和专家搜索。稀疏性也是话题建模不得不考虑的一个问题。北京大学、密歇根大学^[144]提出了利用多类型的上下文的方法将语义集划分为多视图。香港中文大学、密歇根大学^[145]提出了一种双重稀疏话题模型，它不仅在混合话题中强调稀疏性，并且在单词用法上同样重视。通过优先运用一种“Spike and Slab”方法去分离文档 - 话题和话题 - 单词分布的稀疏和平滑，允许个人文档去选择一些关注的话题，同样地，允许一个话题去选择关注的术语。

随着网络信息飞速的发展，收集并组织相关信息变得越来越困难。话题检测与跟踪 (Topic Detection and Tracking, TDT) 就是为解决该问题而提出来的研究方向。话题检测是 TDT 中重要的研究任务之一，其主要研究内容是把讨论相同话题的故事聚类到一起。北京航空航天大学^[146]提出了一种基于增量型聚类的和自动话题检测方法，该方法旨在提高话题检测的效率，并且能够自动检测出文本库中话题的数量。采用改进的权重算法计算特征的权重，通过自适应地提炼具有较强的主题辨别能力的文本特征来提高文档聚类的准确率，并且在聚类过程中利用 BIC 来判断话题类别的数目，同时利用话题的延续性特征来预聚类文档，并以此提高话题检测的速度。话题形态的研究涉及两个问题，其一是话题的结构特性，其二是话题变形。苏州大学^[147]对比分析了现有词包式、层次树式和链式这 3 类主流话题模型的形态特征，尤其深入探讨了静态和动态话题模型拟合话题脉络的优势和劣势，并提出一种基于特征重叠比的核捕捉衰减评价策略，专门用于衡量静态和动态话题模型追踪话题发展趋势的能力。微博具有信息量庞大，信息分散多样等特点，已经成为快速分享和传播信息的新平台。传统话题发现算法大部分都是基于划分的，没有考虑话题之间的关联性，存在一定的局限性，西南交通大学^[148]研究了大规模微博文本集上的话题发现问题。采用具有分词准确率较高、歧义识别特点的西南交通大学思维与智慧研究所中文分词系统对文本进行分词处理，并提出了基于混合模型的微博交叉话题发现算法。

3.6 媒体上下文及标签信息的应用

在媒体上下文及标签信息的应用研究方面，国内学者主要研究了标签相似度、基于用户行为的视频标注、多模态对象相关性、异构媒体语义关系、多媒体文档语义等问题。

中国科技大学、美国 AKiiRA 媒体系统公司和微软亚洲研究院^[149]研究基于标签的社会图像搜索问题，提出了标签相似度的计算方法，将标签形似和视觉相似综合考虑进行图像搜索；香港城市大学和微软亚洲研究院^[150]通过挖掘用户搜索行为，研究包括标签赋值、排序和改进的视频标注；浙江大学^[151]根据不同类型的媒体对象特征及共现信息构建统一的跨媒体关系图，以此统一表达不同模态的媒体对象，通过挖掘不同模态的媒

体对象之间的语义相关性，实现由输入音频实例查询图像结果等此类输入与输出不同模态的跨媒体检索；浙江大学、新加坡南洋理工大学、新加坡国立大学和美国柯达研究院^[152]通过学习异构的多媒体数据的内容，特别是不同媒体类型中的共现信息，获得异构媒体数据的语义关系，构建多媒体关系空间来表达异构多媒体数据，实现跨媒体检索；浙江大学与新加坡南洋理工大学^[153]通过分析异构媒体数据，为一系列带有相同语义的不同媒体类型的对象构建多媒体文档半语义图，建立跨媒体索引空间。

3.7 社会媒体中的内容分析

在社会媒体内容分析方面，国内学者主要研究了基于局部视觉特征的图像搜索、舞蹈音视频搜索、基于内容的音乐搜索、视频内容挖掘等问题。

辽宁师范大学与南京科技大学^[154]提出一种基于 SURF、颜色、显著点等局部视觉特征的图像搜索方法；台湾中正大学^[155]研究如何从舞蹈视频和音乐中抽取节奏信息，根据视频中跳舞者的动作构造运动轨迹并据此判断舞蹈节奏，从音乐中检测拍节来描述音乐的节奏，将这两种模态信息表达为节奏信息序列以支持跨媒体相关性的搜索；台湾大学^[156]开发了一个基于内容的音乐检索系统，对音频资源构建基于听觉特征的索引以及基于标签的索引，根据用户给出的关于多个声道及标注的参数信息，检索相关的音乐；香港城市大学与马来西亚拉曼大学^[157]应用视频内容挖掘和特征选择技术将输入的待查询视频划分成片段，将各片段分别从时间、空间两个不同主线从 Web 上各种知识库中搜索相关的多媒体内容。

3.8 社会媒体搜索技术的产业化

社会媒体搜索技术的研究，在国内同样有相关的产品、系统和专利问世，主要有社会舆情分析与服务系统，面向移动用户的生活信息搜索系统，以及知识图谱系统等。

北京 TRS 公司推出了社会媒体分析云服务^[158]，针对海量社会媒体数据中相关舆情信息的采集、分析与搜索。出门问问是一款面向移动用户的智能语音搜索系统，主要面向的是社交媒体中，如餐饮、娱乐等相关信息的检索^[159]。

2.6 节提及了 Google 公司的知识图谱。实际上，除了 Google 公司之外，国内的百度^[160]和搜狗^[161]也均有自己的知识图谱系统。

在专利申请方面，北京工商大学发明了一种社会化搜索引擎方法，首先基于微博，抽取微博用户的基本信息，建立专家信息库，然后获取用户查询请求，根据查询请求在专家信息库中找到与之相关的专家，利用微博的社会关系提高搜索的准确率^[162]。云壤（北京）信息技术有限公司发明了一个社会化搜索系统，用于搜索至少一个网络社区的多个话题，所述话题包括话题基本信息、话题关联信息、创建该话题的创建成员信息和与该话题关联的关联成员信息^[163]。腾讯科技（深圳）有限公司发明了一种社会化网络中特征关系圈的提取方法，该方法通过提取社会化网络中特征关系图，能够有效利用社

会化网络关系链信息，实现信息有效传播和精确搜索的目标^[164]。中国科学院自动化研究所发明了一种基于多模态生成式模型的自适应社会关系强度挖掘方法，通过收集用户上传的图片信息以及与其有社会关系的用户进行建模，得到一个可以描述用户兴趣分布的主题空间和用户的主题分布，用于自适应的多媒体检索等应用^[165]。

3.9 社会媒体相关科研项目立项

在国家的一些重大研究计划中，将社会媒体搜索相关研究作为重点。近几年国家设立的相关社会媒体的国家重点基础研究发展计划（973 计划）项目有：中科院自动化研究所模式识别国家重点实验室主任谭铁牛研究员主持的“面向公共安全的社会感知数据处理（2012—2016）”项目^[166~170]、清华大学胡世民教授主持的“网络海量可视媒体智能处理的理论与方法（2011—2015）”^[171~174]、浙江大学计算机学院庄越挺教授主持的“面向公共安全的跨媒体计算理论与方法（2012—2016）”项目^[175~177]、北京邮电大学方滨兴院士主持的“社交网络分析与网络信息传播的基础理论研究（2013—2017）”项目^[178~179]、浙江大学大蔡登教授主持的“社交网络信息传播分析与挖掘（2013—2017）”项目^[180~182]。作为热点研究话题，社会媒体搜索和相关研究领域在今年已经获得了多项国家自然科学基金资助。在社会媒体搜索方面，武汉大学的唐晓波开展“社会化媒体集成检索与语义分析方法研究”^[183]；复旦大学的肖仰华开展“面向社会网络的查询处理关键技术研究”^[184]；华东师范大学的王晓玲进行“XML 个性化协作搜索及其在社会网络服务中的应用”研究^[185]；北京航空航天大学的李舟军开展“基于面向话题的加权社会网络的个性化推荐及检索技术研究”^[186]；华中科技大学的陈汉华开展“社交网络搜索系统中基于交互局部性的通信代价优化策略研究”^[187]。

社会媒体中包含各种类别的多媒体信息，因此多媒体分析技术已经成为社会媒体搜索的支撑技术之一。北京大学的崔斌展开“社会化媒体中的数据管理与挖掘研究”^[188]；香港城市大学深圳研究院的杨宗桦开展“面向大规模多媒体检索的异构多模态融合技术研究”^[189]；北京大学彭宇新开展“基于内容的跨媒体检索研究”^[190]。目前已经有一些关于多媒体信息分析技术的重点基金项目获批，包括西安电子科技大学的高新波“多媒体信息处理与分析”^[191]，中国科学院自动化研究所徐常胜的“多媒体内容分析与搜索”^[192]，合肥工业大学汪萌的“多媒体分析与处理”^[193]。

4 国内外研究进展对比

4.1 社会媒体搜索研究现状的特点

纵观社会媒体搜索的研究现状，无论国际还是国内，均呈现出以下主要特点：

1) 复杂图模型是社交媒体用户与资源建模的主要形式。

2) 对象关系抽取是社交媒体资源与用户建模的重要基础，群组探测与话题建模进一步丰富了社交媒体资源与用户模型。

3) 图搜索是社交媒体资源、用户及其关系搜索的理论支撑，群组探测和话题建模为社交媒体搜索提供了更多的搜索条件和更丰富的搜索结果。

4) 媒体上下文及标签信息的应用是社交媒体搜索的新特点。

5) 内容分析仍然是社交媒体搜索不可或缺的技术。

6) 社交媒体搜索技术已在一定程度上产业化。

此外，就广义的社会媒体研究而言，近年来国际顶级会议，如数据库领域的 SIGMOD、VLDB、ICDE，多媒体领域的 ACM MM，信息检索领域的 SIGIR，人工智能领域的 KDD，计算语言学领域的 ACL 等，均有许多学者发表了相关研究论文，其内容涉及特征学习及融合^[115, 194~196]、推荐与搜索^[197~207]、社群发现^[208, 209]、用户兴趣与行为挖掘^[210~216]、事件与话题检测^[217~223]、情感分析^[5, 224]、媒体挖掘^[225~232]等，研究者包括国际、中国港台地区和大陆学者。

4.2 国内外研究现状对比分析

对比国内外研究现状，表现出以下特点：

1) 在理论、方法和技术研究方面，国内在该领域的总体研究水平还相对落后。但是，就国内的领先研究成果而言，如微软亚洲研究院、清华大学、浙江大学等单位的成果，无论研究方法，还是取得成果的质量，并不逊色于国外，也就是说，在该领域中，国内学者具备进行高水平研究的能力。特别是：随着社交媒体的迅速普及，媒体资源共享以及用户群体化、社会化倾向不断增强，并在社会事务中发挥越来越重要的作用，并对现实社会产生了巨大影响。除了上述有关社交媒体的相关研究，国内对相关主题也进行了深入研究。针对多源异构、关系繁杂、持续变化的在线社会关系网络数据，2014 年国内一级学报《计算机学报》以“在线社会关系网络的挖掘和分析”为主题，以专刊发表了有关成果。核心内容包括在线社会网络结构分析^[233~235]，社会网络信息传播^[236~239]，社会网络影响力分析^[240~243]，以及社会媒体网络中朋友推荐^[244]、同行推荐^[245]、统计实证分析^[246]、话题演化^[247]等。2014 年国内一级学报《软件学报》拟以“社会计算：理论与方法”为主题，以专刊发表一期相关研究成果。涉及的内容包括社会计算的基础理论，社交媒体数据感知、社团建模与分析、社会关系分析与挖掘、社会行为分析识别与决策评估等内容，目的是发展社会计算的基础理论、建模、分析与计算方法。

2) 在研究内容方面，由于数据集获取渠道的某些限制，国外研究成果涉及的内容更加广泛，包括 Facebook、YouTub、Flickr、Twitter 等不同形式、不同模态的社交媒体资源，国内研究成果相对狭窄，主要以腾讯、新浪等微博客为主。但是，在以文本为主的社会媒体（如微博、评论等）内容分析中，由于中文的语法、语义分析较英文更加复

杂，国内学者在该领域的研究成果更具特色。

3) 在研究成果的应用方面，虽然国内学者也申请了许多专利、开发了许多系统，但大多数系统尚未有效地投入实际应用。在具有社会媒体搜索功能的搜索引擎中，国内在数量、特别是质量和性能上较国际先进水平均具有相当的差距。

4) 一个值得注意的现象是，大陆学者大多数高水平研究成果均为与港、台地区或者国外学者合作的结果。因此，为尽快提高国内学者的研究水平，寻求国际和地区合作是一个有效的途径和必然的趋势。

5 研究展望

社会媒体中用户与资源关系的搜索问题与传统的搜索系统、社交网络搜索以及社会化搜索具有很大区别。因此，有关社会媒体中用户与资源关系的搜索存在很多问题与挑战。

5.1 问题与挑战

社会媒体搜索不同于传统的多媒体搜索（尽管社会媒体中的信息资源包括各种媒体类型），也不同于关注人与人之间关系的社交网络搜索（尽管社会媒体中同样存在人与人的关系），该搜索着重于社会媒体中资源本身、资源组合、资源与资源之间以及用户与资源之间关系的搜索。针对上述搜索需求及社会媒体的特点，社会媒体中用户与资源及其关系搜索面临如下问题和挑战：

1) 用户与资源异构信息网络建模问题。社会媒体已成为一个巨大的用户与资源构成的复杂关系网络，网络的节点包括用户以及不同模态的媒体资源，节点间的关系既涉及用户对资源的操作方式，也涉及资源之间语义和情境上的关联性，又包括内容的相似性。如何为这样的网络建立一个适用于搜索的模型，是一个新的问题和挑战。

2) 资源间显式与隐式相关性判别问题。社会媒体资源之间的关系包括语义和情境上的关联性，如不同用户发布了来自同一地点或关于同一事件的资源，又包括内容的相似性。如何发现资源之间这些显式和隐式的关联性，是一个新的问题和挑战。同时，由于社会媒体中重复信息、垃圾信息的大量存在，使这一问题更加严重。

3) 社会媒体综合资源发现与组织问题。社会媒体中的单一资源即为原始信息单元，子资源和符合资源通过对单一资源的分割和组合来获得。但是，综合资源并非通过这种简单的方法获得，又是社会媒体搜索中所需要的。因此，如何通过相关资源的分析与挖掘，并结合用户信息获得相关资源并重新组织成综合资源，是一个新的问题与挑战。

4) 基于异构信息网络的资源搜索问题。图搜索本身就是一个尚未很好解决的问题，而社会媒体资源与用户构成的网络图是一个节点类型和模态不同、边的关系及其复杂的异构信息网络，这一问题更加突出。

5) 基于话题的社会媒体社区发现问题。这同样是一个大图中的子图搜索问题，而由于话题随时产生和变化，用户对话题的参与形式亦随时变化，因此涉及话题的发现与追踪、用户参与情况的监控以及检测结果的实时分析。

目前，社交媒体中用户与资源及其关系搜索中的上述问题尚未解决，完善的原型系统亦未见报道。因此，有必要对社交媒体中用户与资源及其关系搜索中的关键技术与科学问题进行深入探讨与研究，并建立一个快速、高效的社会媒体用户与资源搜索技术平台。

5.2 主要研究内容

如前所述，社交媒体搜索有别于传统的搜索系统，以搜索用户关系为主的社交网络搜索，以及较为流行的社会化搜索。因此，对于社会网络搜索的研究，主要包括如下主要研究内容。

(1) 面向社会媒体搜索的资源与用户关系建模

针对社交媒体搜索的需求，将社交媒体抽象成用户和资源及其关系。为合理地描述用户与用户、用户与资源、资源与资源之间的关系，并支持高效的搜索，需要研究一种合理的模型框架。目前，复杂的图数据结构是构建这一模型的理想选择，为此，需要研究具体的图结构、模式及相关的存储及索引机制。

(2) 基于情境感知的社会媒体资源关联分析

将资源、用户均视为实体，研究实体识别、实体消歧、重复与无效实体过滤。实体作为社会媒体资源与用户关系模型（图结构）的顶点，研究实体之间的关系，结果作为模型的边。在分析实体间关系时，将研究跨媒体、不同模态实体的相似性度量，媒体上下文线索、标签、情境、语义信息在实体相关性分析中的应用，最终发现实体之间的显式与隐式关系，并赋予不同的权重。

(3) 基于深度内容分析的社会媒体资源挖掘

综合资源不是社交媒体中现成的实体，而是通过对实体内容、结构的深度分析与挖掘，并重新组织后获得的。因此，将研究与多模态资源分析相关的各种多媒体数据挖掘方法，并结合用户信息及其对资源的操作信息进行重新组织。

(4) 多（跨）模态社会媒体资源及其关系搜索

基于社会媒体资源与用户关系模型，研究复杂图结构中的顶点及子图搜索问题，将涉及图的搜索策略、图的相似性匹配。还将研究搜索结果的排序问题，为此，当搜索者本身即为社交媒体用户时，将结合社会媒体资源与用户关系模型中用户的相关信息。还将研究搜索过程中用户对结果的操作及其向搜索系统的反馈机制。

(5) 基于用户与资源关系的社会媒体社区搜索

社交媒体中，用户通过对资源的操作（如针对某个话题的讨论）而形成社区，为此，将研究新话题的检测及基于该话题的用户社区发现，涉及异构图上的各种数据挖掘技术。

6 总结

社会媒体搜索不同于已有传统的搜索，为此，呈现出了一些新的研究问题和挑战。社会媒体搜索是关注社会媒体中用户与资源、资源与资源关系的搜索，其研究成果将具有重要的学术价值和广阔的应用领域：

1) 支持社会媒体搜索相关技术的研究。从研究的角度，提出的问题与挑战以及研究内容将涉及社会媒体用户与资源的建模、社会媒体信息的挖掘、跨媒体与多模态的资源搜索、资源关系搜索、资源与用户关系搜索等理论与技术的研究，其研究结果可以支持社会媒体搜索相关技术的研究。

2) 支持社会媒体搜索系统的开发。从应用的角度，如果开发者获得足够的社会媒体资源，在实现了输入接口的基础上，应用相应的研究结果，有望开发出具有较完整功能的社会媒体搜索系统。

3) 改进特定网站或需求的搜索质量。对于功能相对单一的社会媒体网站（比如在线购物、在线评论等），用户访问历史、网站媒体资源、用户接口等要素均完备，甚至原本就具有搜索功能，应用相关研究成果，更有望改进搜索质量和用户体验。

4) 提升企事业单位的管理水平。可将相关研究成果融入单位的办公业务模型，应用相关研究成果可获取有关企事业单位发展的多维度的媒体资源，基于社会媒体的综合搜索结果可有效提升领导决策的准确性，有利于运营最佳的客户关系管理模式，有助于提升企事业单位的科学化管理水平。

此外，由于社会媒体用户广泛参与、由用户生成内容构成的，因此，在很大程度上反映出社情民意及社会动向。社会媒体用户与资源关系搜索技术的研究，可为政府职能部门跟踪和分析网络舆情（例如对敏感信息资源的发现及与之相关用户的搜索）提供技术支撑。因此，有关社会媒体搜索的研究具有重要的社会意义。

参考文献

- [1] 杨兴建. 社会化媒体与搜索营销, <http://wenku.baidu.com/view/ee02d983bceb19e8b8f-6badf.html>.
- [2] 张辉等. 速途研究院: 2013 年社会化媒体分析报告. <http://www.sootoo.com/content/464520.shtml#>.
- [3] 黄立威, 李德毅. 社交媒体中的信息推荐 [J]. 智能系统学报. 2012, 7(1): 1-8.
- [4] 社会化搜索, <http://baike.baidu.com/view/427040.htm?qq-pf-to=pcqq.c2c>.
- [5] Zhu L, Galstyan A, Cheng J, Lerman K. Tripartite graph clustering for dynamic sentiment analysis on social media. SIGMOD 2014: 1531-1542.
- [6] Kao A, Ferng W, Poteet S, Quach L, Tjoelker R. TALISON- Tensor analysis of social media data. ISI 2013: 137-142.
- [7] Kim M, Newth D, Christen P. Modeling dynamics of meta- populations with a probabilistic approach;

- global diffusion in social media. CIKM 2013: 489-498.
- [8] Chelmis C, Prasanna V. Exploring generative models of tripartite graphs for recommendation in social media. MSM 2013: 2.
- [9] Ahmed A, Hong L, Smola A. Hierarchical geographical modeling of user locations from social media posts. WWW 2013: 25-36.
- [10] Valera I, Gomez-Rodriguez M, Gummadi K. Modeling Diffusion of Competing Products and Conventions in Social Media. CoRR abs/1406.0516(2014).
- [11] Liu D, Blenn N, Mieghem P. Modeling Social Networks with Overlapping Communities Using Hypergraphs and Their Line Graphs CoRR abs/1012.2774(2010).
- [12] Lau R, Xia Y, Ye Y. A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media. IEEE Comp. Int. Mag. (CIM)9(1): 31-43(2014).
- [13] Elwood S, McCaleb K, Fernandez M, Keengwe J. A theoretical framework and model towards media-rich social presence design practices. EAIT 19(1): 239-249(2014).
- [14] Darmon D, Sylvester J, Girvan M, Rand W. Predictability of User Behavior in Social Media: Bottom-Up v. Top-Down Modeling. SocialCom 2013: 102-107.
- [15] Kim Mg, Newth D, Christen P. Modeling Dynamics of Diffusion Across Heterogeneous Social Networks: News Diffusion in Social Media. Entropy 15(10): 4215-4242(2013).
- [16] Woo J, Son J, Chen H. An SIR model for violent topic diffusion in social media. ISI 2011: 15-19.
- [17] Vavliakis K, Symeonidis A, Mitkas P. Event identification in web social media through named entity recognition and topic modeling. Data Knowl. Eng. (DKE)88: 1-24(2013).
- [18] Yerva S, Catasta M, Demartini G, Aberer K. Entity disambiguation in tweets leveraging user social profiles. IRI 2013: 120-128.
- [19] Gattani A, Lamba D, Garera N, et al. . Entity Extraction, Linking, Classification, and Tagging for Social Media: A Wikipedia-Based Approach. PVLDB 6(11): 1126-1137(2013).
- [20] Wu P, Hoi S, Xia H, Zhao P, et al. . Online multimodal deep similarity learning with application to image retrieval. ACM Multimedia 2013: 153-162.
- [21] Papaoikonomou A, Tserpes K, Kardara M, Varvarigou T. A Similarity-based Chinese Restaurant Process for Social Event Detection. MediaEval 2013.
- [22] Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. Effects of user similarity in social media. WSDM 2012: 703-712.
- [23] Beecks C, Kirchhoff S, Seidl T. On stability of signature-based similarity measures for content-based image retrieval. Multimedia Tools Appl. (MTA)71(1): 349-362(2014).
- [24] Lee M, Lee K, Park J. Music similarity-based approach to generating dance motion sequence. Multimedia Tools Appl. (MTA)62(3): 895-912(2013).
- [25] Wolf L, Levy N. The SVM-Minus Similarity Score for Video Face Recognition. CVPR 2013: 3523-3530.
- [26] Lee S, Sim J, Kim C, Lee S. Correspondence Matching of Multi- View Video Sequences Using Mutual Information Based Similarity Measure. IEEE Transactions on Multimedia(TMM)15(8): 1719-1731(2013).
- [27] Curtiss M, Becker I, Bosman T, et al. . Unicorn: A System for Searching the Social Graph. PVLDB 6(11): 1150-1161(2013).
- [28] Sulieman D, Malek M, Kadima H, Laurent D. Graph Searching Algorithms for Semantic-Social Recommendation. ASONAM 2012: 733-738.

- [29] Veloso R, Cerf L, Junior W, Zaki M. Reachability Queries in Very Large Graphs: A Fast Refined Online Search Approach. EDBT 2014: 511-522.
- [30] Flerova N, Marinescu R, Dechter R. Weighted Best First Search for Graphical Models. ISAIM 2014.
- [31] Sun Y, Han J. Mining Heterogeneous Information Networks: Principles and Methodologies Morgan & Claypool Publishers 2012.
- [32] Yu X, Sun Y, Norick B, Mao T, Han J. User guided entity similarity search using meta-path selection in heterogeneous information networks. CIKM 2012: 2025-2029.
- [33] Sun Y, Han J, Yan X, Yu P, Wu T. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. PVLDB 4(11): 992-1003(2011).
- [34] He J, Bailey J, Zhang R. Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks. DASFAA 2014: 141-155.
- [35] Shi Jianbo, Malik Jitendra. Normalized Cuts and Image Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888 ~ 905.
- [36] Newman M E J. Modularity and Community Structure in Networks. Proceedings of the National Academy of Science, 2006, 103(23): 8577 ~ 8582.
- [37] Chang Yu-Teng, Leahy Richard M, Pantazis Dimitrios. Modularity-Based Graph Partitioning Using Conditional Expected Models. Physics Review E, 2012, 85(1): 16109.
- [38] Chakrabarti Deepayan. Autopart. Parameter-Free Graph Partitioning and Outlier Detection[C]. In: Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases. New York, NY, USA: Springer-Verlag, 2004. 112 ~ 124.
- [39] Sun Jimeng, Papadimitriou Spiros, Yu Philip S, et al. Graphscope: Parameter-Free Mining of Large Time-Evolving Graphs [C]. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007. 687 ~ 696.
- [40] Rosvall Martin, Bergstrom Carl T. An Information-Theoretic Framework for Resolving Community Structure in Complex Networks. Proceedings of the National Academy of Sciences, 2007, 104(18): 7327 ~ 7331.
- [41] Dhillon Inderjit S, Mallela Subramanyam, Modha Dharmendra S. Information-Theoretic Co-Clustering[C]. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2003. 89 ~ 98.
- [42] Luxburg Ulrike. A Tutorial on Spectral Clustering. Statistics and Computing, 2007, 17(4): 395 ~ 416.
- [43] Hartigan John A, Wong M Anthony. A K-Means Clustering Algorithm. Applied Statistics, 1979, 28(1): 100 ~ 108.
- [44] Nowicki Krzysztof, Snijders Tom A B. Estimation and Prediction for Stochastic Blockstructures[J]. Journal of the American Statistical Association, 2001, 96(455): 1077 ~ 1087.
- [45] Airolodi Edoardo M, Blei David M, Fienberg Stephen E, et al. Mixed Membership Stochastic Blockmodels [J]. The Journal of Machine Learning Research, 2008, 9(6): 1981 ~ 2014.
- [46] Chakrabarti Deepayan, Kumar Ravi, Tomkins Andrew. Evolutionary Clustering[C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2006. 554 ~ 560.
- [47] Chi Yun, Song Xiaodan, Zhou Dengyong, et al. Evolutionary Spectral Clustering by Incorporating Temporal Smoothness[C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2007. 153 ~ 162.

- [48] Lin Yu-Ru, Chi Yun, Zhu Shenghuo, et al. Analyzing Communities and their Evolutions in Dynamic Social Networks. *ACM Transactions on Knowledge Discovery from Data*, 2009, 3(2) : 1 ~ 31.
- [49] Kim Min-Soo, Han Jiawei. A Particle- and-Density Based Evolutionary Clustering Method for Dynamic Networks. *Proceedings of the VLDB Endowment*, 2009, 2(1) : 622 ~ 633.
- [50] Yang Tianbao, Chi Yun, Zhu Shenghuo, et al. A Bayesian Approach Toward Finding Communities and their Evolutions in Dynamic Social Networks[C]. In: *Proceedings of the SIAM International Conference on Data Mining*. Philadelphia, PA, USA: SIAM, 2009. 990 ~ 1001.
- [51] Chang-Dong Wang, Jian-Huang Lai, Philip S. Yu. Dynamic Community Detection in Weighted Graph Streams[C]. *Proceedings of the 2013 SIAM International Conference on Data Mining*. 2013. 151 ~ 161.
- [52] Karthik Subbian, Charu C. Aggarwal, Jaideep Srivastava, Philip S. Yu. Community Detection with Prior Knowledge[C]. In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. 2013. 405 ~ 413.
- [53] G Palla, A L Barabasi, T Vicsek. Quantifying Social Group Evolution. *Nature*, 2007, 446(7136) : 664 ~ 667.
- [54] Asur Sitaram, Parthasarathy Srinivasan, Ucar Duygu. An Event-Based Framework for Characterizing the Evolutionary Behavior of Interaction Graphs. *ACM Transactions on Knowledge Discovery from Data*, 2009, 3(4) : 1 ~ 36.
- [55] Greene Derek, Doyle Donal, Cunningham Padraig. Tracking the Evolution of Communities in Dynamic Social Networks [C]. In: *International Conference on Advances in Social Networks Analysis and Mining*. New York, NY, USA: ACM, 2010. 176 ~ 183.
- [56] Tantipathananandh Chayant, Berger-Wolf Tanya Y, Kempe David. A Framework for Community Identification in Dynamic Social Networks[C]. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2007. 717 ~ 726.
- [57] Lin Yu-Ru, Sun Jimeng, Castro Paul, et al. MetaFac: Community Discovery via Relational Hypergraph Factorization[C]. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2009. 527 ~ 536.
- [58] Tang Lei, Liu Huan, Zhang Jianping, et al. Community Evolution in Dynamic Multi-Mode Networks[C]. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. 677 ~ 685.
- [59] Sun Yizhou, Tang Jie, Han Jiawei, et al. Community Evolution Detection in Dynamic Heterogeneous Information Networks. In: *Proceedings of the 8th Workshop on Mining and Learning with Graphs*. New York, NY, USA: ACM, 2010. 137 ~ 146.
- [60] Ning Huazhong, Xu Wei, Chi Yun, et al. Incremental Spectral Clustering by Efficiently Updating the Eigen-System. *Pattern Recognition*, 2010, 43(1) : 113 ~ 127.
- [61] Hofmann, T. Probabilistic Latent Semantic Analysis [C]. in *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. 1999. Arlington, Virginia, USA: AUAI.
- [62] Blei D M, A Y Ng, M I Jordan. Latent dirichlet allocation[J]. *Journal of Machine Learning Research*, 2003. 3(1) : 993-1022.
- [63] Heinrich G. Parameter estimation for text analysis. 2008, University of Leipzig.
- [64] Andrzejewski David, Buttler David. Latent Topic Feedback for Information Retrieval[C]. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2011. 600 ~ 608.
- [65] Mei Q, et al. Topic modeling with network regularization[C]. In *Proceedings of the 17th international*

- conference on World Wide Web. 2008. New York, NY, USA: ACM.
- [66] Sun Y, et al. iTopicModel: Information Network-Integrated Topic Modeling[C]. In Proceedings of the 9th IEEE International Conference on Data Mining. 2009. Washinton, DC, USA: IEEE.
- [67] Lin C X, et al. PET: a statistical model for popular events tracking in social communities [C]. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010. New York, NY, USA: ACM.
- [68] Deng H, et al. Probabilistic topic models with biased propagation on heterogeneous information networks [C]. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011. New York, NY, USA: ACM.
- [69] Chang J, D M Blei. Relational Topic Models for Document Networks[J]. Journal of Machine Learning Research-Proceedings Track , 2009. 5(1) : 81-88.
- [70] Nallapati R, D A McFarland, C D Manning. TopicFlow Model: Unsupervised Learning of Topic-specific Influences of Hyperlinked Documents [J]. Journal of Machine Learning Research- Proceedings Track , 2011. 15(1) : p. 543-551.
- [71] Blei D M, J D Lafferty. Dynamic topic models[C]. In Proceedings of the 23rd International Conference on Machine Learning. 2006. New York, NY, USA: ACM.
- [72] Wang C, D M Blei, D Heckerman. Continuous Time Dynamic Topic Models[C]. In Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence. 2008. Arlington, Virginia, USA: AUAI Press.
- [73] Pruteanu-Malinici I, et al. Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents. IEEE Transactions on Pattern Analysis and Machine Intelligence , 2010. 32(6) : p. 996-1011.
- [74] Ahmed A, E P Xing. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream [C]. In UAIProceedings of the 26th Conference on Uncertainty in Artificial Intelligence. 2010. Arlington, Virginia, USA: AUAI.
- [75] Iwata T, et al. Online multiscale dynamic topic models[C]. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2010. New York, NY, USA: ACM.
- [76] Wang X, A McCallum. Topics over time: a non-Markov continuous-time model of topical trends[C]. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining. 2006. New York, NY, USA: ACM.
- [77] Saha A, V Sindhwani. Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization[C]. In Proceedings of the 5th ACM International Conference on Web Search and Data Mining. 2012. New York, NY, USA: ACM.
- [78] Blei D M, J D McAuliffe. Supervised Topic Models[C]. In Proceedings of the 21st Annual Conference on Neural Information Processing Systems. 2007. Cambridge, MA, USA: MIT Press.
- [79] Wang C, D M Blei. Collaborative topic modeling for recommending scientific articles[C]. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011. New York, NY, USA: ACM.
- [80] Hong L, et al. Tracking trends: incorporating term volume into temporal topic models[C]. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2011. New York, NY, USA: ACM.
- [81] Ahmed A, et al. Scalable distributed inference of dynamic user interests for behavioral targeting[C]. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data

- Mining. 2011. New York, NY, USA: ACM.
- [82] Balasubramanyan R, et al. Modeling Polarizing Topics When Do Different Political Communities Respond Differently to the Same News [C]. In Proceedings of the 6th International Conference on Weblogs and Social Media. 2012. Menlo Park, CA, USA: AAAI.
- [83] Gerrish S, D M Blei. Predicting Legislative Roll Calls from Text [C]. In Proceedings of the 28th International Conference on Machine Learning. 2011. Madison, WI, USA: Omnipress.
- [84] Saxton G, Oh O, Kishore R. Rules of Crowdsourcing: Models, Issues, and Systems of Control. IS Management (ISM)30(1) : 2-20(2013).
- [85] Rawashdeh M, Kim H, El-Saddik A. Social Media Annotation and Tagging Based on Folksonomy Link Prediction in a Tripartite Graph. MMM 2013 : 24-35.
- [86] Kim H, Rawashdeh M, Alghamdi A, El-Saddik A. Folksonomy-based personalized search and ranking in social media services. Inf. Syst. (IS)37(1) : 61-76(2012).
- [87] Rawashdeh M, Kim Hm, El-Saddik A. Folksonomy-boosted social media search and ranking. ICMR 2011 : 27.
- [88] Clements M, Vries A, Reinders M. The influence of personalization on tag query length in social media search. Inf. Process. Manage. (IPM)46(4) : 403-412(2010).
- [89] Saari P, Eerola T. Semantic Computing of Moods Based on Tags in Social Media of Music. CoRR abs/1308.1817(2013).
- [90] Kiciman E, Counts S, Gamon M, Choudhury M, Thiesson B. Discussion Graphs: Putting Social Media Analysis in Context. ICWSM 2014.
- [91] Jiang L, Yang C. Using Co-occurrence Analysis to Expand Consumer Health Vocabularies from Social Media Data. ICHI 2013 : 74-81.
- [92] Dereczynski L, Yang B, Jensen C. Towards context-aware search and analysis on social media data. EDBT 2013 : 137-142.
- [93] Zhang H, Zha Z, Yang Y, Yan S, Gao Y, Chua T. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. ACM Multimedia 2013 : 33-42.
- [94] Burdescu D, Brezovan M, Ganea E, Stanescu L. New Algorithm for Segmentation of Images Represented as Hypergraph Hexagonal-Grid. IbPRIA 2011 : 395-402.
- [95] Ding L, Yilmaz A. Interactive image segmentation using probabilistic hypergraphs. Pattern Recognition (PR) 43(5) : 1863-1873(2010).
- [96] Liu L, Xu H, Xing J, Liu S, Zhou X, Yan S. Wow! you are so beautiful today! . ACM Multimedia 2013 : 3-12.
- [97] ElAlami M. A new matching strategy for content based image retrieval system. Appl. Soft Comput. (ASC)14 : 407-418(2014).
- [98] Seetharaman K, Kamarasan M. Statistical framework for content-based medical image retrieval based on wavelet orthogonal polynomial model with multiresolution structure. IJMIR 3(1) : 53-66(2014).
- [99] Chattopadhyay C, Maurya A. Multivariate time series modeling of geometric features of spatio-temporal volumes for content based video retrieval. IJMIR 3(1) : 15-28(2014).
- [100] Chattopadhyay C, Maurya A. Genre-specific modeling of visual features for efficient content based video shot classification and retrieval. IJMIR 2(4) : 289-297(2013).
- [101] Xie L, Shamma D, Snoek C. Content is dead; long-live content! [C]. ACM Multimedia. Nara, Japan,

- 2012: 7-8.
- [102] <https://www.facebook.com/about/graphsearch>.
- [103] Wable A, DeLorme L, Kao W, Roche A, Occhino T. Search and retrieval of objects in a social networking system. American Patents. 2012.
- [104] <http://www.google.com/insidesearch/features/search/knowledge.html>.
- [105] <https://twitter.com/search-home>.
- [106] Gilbert T. Social media content aggregation and search mechanism. American Patents. 2013.
- [107] Codella N, Natsev A, John R. Social media event detection and content-Based retrieval. American Patents. 2014.
- [108] Yao T, Liu Y, Ngo C, Mei T. Unified entity search in social media community. WWW 2013: 1457-1466.
- [109] Zhuang Y, Liu Y, Wu F, Zhang Y, Shao J. Hypergraph spectral hashing for similarity search of social image. ACM Multimedia 2011: 1457-1460.
- [110] Gao Y, Wang M, Luan H, Shen J, Yan S, Tao D. Tag-based social image search with visual-text joint hypergraph learning. ACM Multimedia 2011: 1517-1520.
- [111] C. Hu, C. Zhang, T. Wang, Q. Li. An Adaptive Recommendation System in Social Media. HICSS 2012: 1759-1767.
- [112] Wang D, Liu X, Luo H, Fan J. Semantic Entity Identification in Large Scale Data via Statistical Features and DT-SVM. WISE 2013: 354-367.
- [113] Liu L, Zhu F, Jiang M, Han J, Sun L, Yang S. Mining Diversity on Social Media Networks. Multimedia Tools Appl. (MTA)56(1): 179-205(2012).
- [114] Ling L, Zhai X, Peng Y. Tri-Space and Ranking Based Heterogeneous Similarity Measure for Cross-Media Retrieval. ICPR 2012: 230-233.
- [115] Cui B, Tung A, Zhang C, Zhao Z. Multiple feature fusion for social media applications. SIGMOD 2010: 435-446.
- [116] Zhu X, Huang Z, Shen H, Zhao X. Linear Cross-Modal Hashing for Efficient Multimedia Search. ACM Multimedia 2013: 143-152.
- [117] Wu S, Tang J, Gao B. Instant Social Graph Search. PAKDD 2012: 256-267.
- [118] Lu B, Yuan Y, Wang G. SRGSIS: a novel framework based on social relationship graph for social image search. CIKM 2012: 2615-2618.
- [119] Chiang M, Liou J, J Wang J, Peng W, Shan M. Exploring heterogeneous information networks and random walk with restart for academic search. Knowl. Inf. Syst. (KAIS)36(1): 59-82(2013).
- [120] Tan S, Bu J, Qin X, Chen C, Cai D. Cross domain recommendation based on multi-type media fusion. Neurocomputing (IJON)127: 124-134(2014).
- [121] Chen L, Zeng W, Yuan Q. A unified framework for recommending items, groups and friends in social media environment via mutual resource fusion. Expert Syst. Appl. (ESWA)40(8): 2889-2903(2013).
- [122] 汪小帆, 刘亚冰. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报, 2009, 38(5): 537 ~ 543.
- [123] 程学旗, 沈华伟. 复杂网络的社区结构[J]. 复杂系统与复杂性科学, 2011, 8(1): 57 ~ 70
- [124] 沈华伟, 程学旗, 陈海强等. 基于信息瓶颈的社区发现[J]. 计算机学报, 2008, 31(4): 677 ~ 686.
- [125] Liang Huang, Ruixuan Li, Hong Chen, et al. Detecting Network Communities Using Regularized Spectral Clustering Algorithm. Artificial Intelligence Review, 2012: 1 ~ 16.

- [126] 李兆南, 杨博, 刘大有. 复杂网络社区挖掘的距离相似度算法[J]. 计算机科学与探索. 2011; 5: 336.
- [127] 方平, 郭正彪, 李芝棠等. 基于共同好友数的在线社会网络社区发现算法[J]. 计算机科学与探索. 2012; 6: 456.
- [128] 黄发良, 张师超, 朱晓峰. 基于多目标优化的网络社区发现方法[J]. 软件学报. 2014-01; 24: 2062-2077.
- [129] Duan Dongsheng, Li Yuhua, Jin Yanan, et al. Community Mining on Dynamic Weighted Directed Graphs. In: Proceedings of the ACM 1st International Workshop on Complex Networks Meet Information & Knowledge Management (CIKM-CNIKM). New York, NY, USA: ACM, 2009. 11 ~ 18.
- [130] Duan Dongsheng, Li Yuhua, Li Ruixuan, et al. Incremental K-Clique Clustering in Dynamic Social Networks. Artificial Intelligence Review, 2012, 38(2): 129 ~ 147.
- [131] 索勃, 李战怀, 陈群, 王忠. 基于信息流动分析的动态社区发现方法[J]. 软件学报. 2014, 25 (3): 547 ~ 559.
- [132] 张春英, 郭景峰. 集对社会网络 α 关系社区及动态挖掘算法[J]. 计算机学报. 2014, 36 (3): 1682-1692.
- [133] 林旺群, 邓镭, 丁兆云, 贾焰, 周斌. 一种新型的层次化动态社区并行计算方法[J]. 计算机学报. 2012, 8: 1712.
- [134] 柴变芳, 于剑, 贾彩燕, 王静红. 一种基于随机块模型的快速广义社区发现算法[J]. 软件学报. 2013, 24(11): 2699 2709.
- [135] 徐永成, 陈峻. 基于蚁群优化的二分网络社区挖掘[J]. 计算机科学与探索. 2014, 297-304.
- [136] 沈华伟等. 基于信息瓶颈的社区发现[J]. 计算机学报. 2008, 31: 677.
- [137] 林旺群, 卢风顺, 丁兆云, 吴泉源, 周斌, 贾焰. 基于带权图的层次化社区并行计算方法[J]. 软件学报. 2012, 23(6): 1517 ~ 1530.
- [138] 燕飞, 张铭, 谭裕韦, 等. 综合社会行动者兴趣和网络拓扑的社区发现方法[J]. 计算机研究与发展, 2010, 47(z1): 357 ~ 362.
- [139] Wang Liaoruo, Lou Tiancheng, Tang Jie et al. Detecting Community Kernels in Large Social Networks [C]. In: Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM 2011), Washington, DC, USA: IEEE, 2011. 784 ~ 793.
- [140] Dongsheng Duan, Yuhua Li, Ruixuan Li, Rui Zhang, Aiming Wen. RankTopic: Ranking Based Topic Modeling. IEEE International Conference on Data Mining (ICDM 2012), Brussels, Belgium, December 10-13, 2012, 211-220.
- [141] Dongsheng Duan, Yuhua Li, Ruixuan Li, Zhengding Lu, Aiming Wen [J]. MEI: Mutual Enhanced Infinite Community-Topic Model for Analyzing Text-Augmented Social Networks. The Computer Journal, 56(3): 336-354(2013), 2013.
- [142] 吴良, 黄威靖, 陈薇等. ACT-LDA: 集成话题, 社区和影响力分析的概率模型[J]. 计算机科学与探索, 2013, 7(8): 718-728.
- [143] Tang J, et al. ArnetMiner: Extraction and Mining of Academic Social Networks[J]. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008. New York, NY, USA: ACM.
- [144] Tang J, Zhang M, Mei Q. One theme in all views: modeling consensus topics in multiple contexts[C]. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data

- mining. ACM, 2013: 5-13.
- [145] Lin T, Tian W, Mei Q, et al. The dual-sparse topic model: mining focused topics and focused terms in short text[C]. In Proceedings of the 23rd international conference on World wide web. International World Wide Web Conferences Steering Committee, 2014: 539-550.
- [146] 张小明, 李舟军, 巢文涵. 基于增量型聚类的自动话题检测研究[J]. 软件学报, 2012, 23(6): 1578-1587.
- [147] 洪宇, 仓玉, 姚建民, 等. 话题跟踪中静态和动态话题模型的核捕捉衰减[J]. 软件学报, 2012, 23(5): 1100-1119.
- [148] 詹勇, 杨燕, 王红军. 混合模型的微博交叉话题发现[J]. 计算机科学与探索, 2013, 7(8): 747-753.
- [149] Yang K, Wang M, Hua X, Zhang H. Tag-Based Social Image Search: Toward Relevant and Diverse Results. Social Media Modeling and Computing 2011: 25-45.
- [150] Yao T, Mei T, Ngo C, Li S. Annotation for free: video tagging by mining user search behavior. ACM Multimedia 2013: 977-986.
- [151] Zhuang Y, Yang Y, Wu F. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Transactions on Multimedia(TMM)10(2): 221-229(2008).
- [152] Yang Y, Xu D, Nie F, Luo J, Zhuang Y. Ranking with local regression and global alignment for cross media retrieval. ACM Multimedia 2009: 175-184.
- [153] Yang Y, Wu F, Xu D, Zhuang Y, Chia L. Cross-media retrieval using query dependent search methods. Pattern Recognition(PR)43(8): 2927-2936(2010).
- [154] Yang H, Li Y, Li W, Wang X, Yang F. Content-based image retrieval using local visual attention feature. J. Visual Communication and Image Representation(JVCIR)25(6): 1308-1323(2014).
- [155] Chu W, Tsai S. Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos. IEEE Transactions on Multimedia(TMM)2012, 14(1): 129-141.
- [156] Wang J, Wu M, Wang H, Jeng S. Query by multi-tags with multi-level preferences for content-based music retrieval. ICME 2011: 1-6.
- [157] Tan S, Ngo C, Tan H, Pang L. Cross media hyperlinking for search topic browsing. ACM Multimedia 2011: 243-252.
- [158] <http://www.trs.com.cn/product/product-smas.html>.
- [159] <http://chumenwenwen.net/>.
- [160] <http://baike.baidu.com/view/10608644.htm>.
- [161] <http://www.sogou.com/labs/webservice/tupu.html>.
- [162] 王恺, 莫倩, 张树, 张传文, 李阳. 一种社会化的搜索引擎方法和系统. 北京工商大学, 中国专利, 2013.
- [163] 刘骏, 孙峥, 盛佳, 李大海, 王东, 陈利人, 曲径, 项锟, 安兴华, 马俊, 寇黎钦, 马剑, 张晓鑫. 社会化搜索系统及搜索方法. 云壤(北京)信息技术有限公司, 中国专利, 2013.
- [164] 蔡耿平, 胡海斌. 一种社会化网络中特征关系圈的提取方法及装置. 腾讯科技(深圳)有限公司, 中国专利, 2009.
- [165] 徐常胜, 桑基韬. 一种基于多模态自适应社会关系强度挖掘的社会搜索方法. 中国科学院自动化研究所, 中国专利, 2013.
- [166] Chen Y, Wang L, Wang W, Zhang Z. Continuum regression for cross-modal multimedia retrieval. ICIP

2012: 1949-1952.

- [167] Zhou Z, Wang W, Wang L. Community Detection Based on an Improved Modularity. CCPR 2012: 1949-1952.
- [168] Zhao F, Huang Y, Wang L, Tan T. Relevance Topic Model for Unstructured Social Group Activity Recognition. NIPS 2013: 2580-2588.
- [169] Zhou G, Liu K, Zhao J. Joint Relevance and Answer Quality Learning for Question Routing in Community QA. CIKM 2012: 1492-1496.
- [170] Zhang T, Liu K, Zhao J. Cross Lingual Entity Linking with Bilingual Topic Model. IJCAI 2013: 2218-2224.
- [171] Yuan Y, Wang G, Chen L, Wang H. Efficient Keyword Search on Uncertain Graph Data. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(12): 2767-2779.
- [172] Zheng Y, Chen P. Clustering based on enhanced α -expansion move. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(10): 2206-2216.
- [173] Zhang F, Wang M, Hu S. Aesthetic Image Enhancement by Dependence-Aware Object Re-Composition. IEEE Transactions on Multimedia, 2013, 15(7): 1480-1490.
- [174] Zhang S, Li X, Hu S, Martin R. Online Video Stream Abstraction and Stylization. IEEE Transactions on Multimedia, 2011, 13(6): 1286-1294.
- [175] Zhang Y, Li G, Chu L, Wang S, Zhang W, Huang Q. Cross-media topic detection: A multi-modality fusion framework. ICME 2013 : 1-6.
- [176] Gao Y, Wang M, Zha Z, Shen J, Li X, Wu X. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. IEEE Transactions on Image Processing, 2013, 22(1): 363-376.
- [177] Gao H, Tang S, Zhang Y, Jiang D, Wu F, Zhuang Y. Supervised Cross-collection Topic Modeling. MM 2012: 957-960.
- [178] Zhang L, Jia Y, Zhou B, Han Y. Detecting real-time burst topics in microblog streams: how sentiment can help. WWW 2013: 781-782.
- [179] Han Y, Fang B, Jia Y. Predicting the Topic Influence Trends in Social Media with Multiple models. Neurocomputing, in Press.
- [180] Wang B, Wang C, Bu J, Chen C, Zhang W, Cai D, He X. Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems. WWW 2013: 1331-1340.
- [181] Mao X, Lin B, Cai D, He X, Pei J. Parallel field alignment for cross media retrieval. MM 2013: 897-906.
- [182] Ziyu Guan, Gengxin Miao, Russell McLoughlin, Xifeng Yan, Deng Cai. Co-Occurrence-Based Diffusion for Expert Search on the Web. IEEE Transactions on Knowledge and Data Engineering, 2013, 25 (5): 1001-1014.
- [183] 唐晓波. 社会化媒体集成检索与语义分析方法研究. 武汉大学, 2013. 1-2016. 12.
- [184] 肖仰华. 面向社会网络的查询处理关键技术研究. 复旦大学, 2011. 1-2013. 12.
- [185] 王晓玲. XML个性化协作搜索及其在社会网络服务中的应用. 华东师范大学, 2012. 1-2014. 12.
- [186] 李舟军. 基于面向话题的加权社会网络的个性化推荐及检索技术研究. 北京航空航天大学, 2012. 1-2014. 12.
- [187] 陈汉华. 社交网络搜索系统中基于交互局部性的通信代价优化策略研究. 华中科技大学, 2014. 1-2017. 12.

- [188] 崔斌. 社会化媒体中的数据管理与挖掘研究. 北京大学, 2011. 1-2013. 12.
- [189] 杨宗桦. 面向大规模多媒体检索的异构多模态融合技术研究. 香港城市大学深圳研究院, 2013. 1-2016. 12.
- [190] 彭宇新. 基于内容的跨媒体检索研究. 北京大学, 2014. 1-2017. 12.
- [191] 高新波. 多媒体信息处理与分析. 西安电子科技大学, 2012. 1-2015. 12.
- [192] 徐常胜. 多媒体内容分析与搜索. 中国科学院自动化研究所, 2013. 1-2016. 12.
- [193] 汪萌. 多媒体分析与处理. 合肥工业大学, 2014. 1-2017. 12.
- [194] Yuan Z, Sang J, Liu Y, Xu C. Latent feature learning in social media network. ACM Multimedia 2013: 253-262.
- [195] Roy S, Mei T, Zeng W, Li S. SocialTransfer: cross-domain transfer learning from social streams for media applications. ACM Multimedia 2012: 649-658.
- [196] Tang J, Liu H. Unsupervised feature selection for linked social media data. KDD 2012: 904-912.
- [197] Liu D, Ye G, Chen C, Yan S, Chang S. Hybrid social media network. ACM Multimedia 2012: 659-668.
- [198] Sang J. Collective search and recommendation in social media. ACM Multimedia 2012: 1421-1424.
- [199] Tan S, Chen C, Wang C, Wu H, Zhang L, He X. Music recommendation by unified hypergraph: combining social media information and music content. ACM Multimedia 2010: 391-400.
- [200] Gupta S, Phung D, Adams B, Tran T, Venkatesh S. Nonnegative shared subspace learning and its application to social media retrieval. KDD 2010: 1169-1178.
- [201] Ronen I, Guy I, Kravi E, Barnea M. Recommending social media content to community owners. SIGIR 2014: 243-252.
- [202] Wan S, Lan Y, Guo J, Fan C, Cheng X. Informational friend recommendation in social media. SIGIR 2013: 1045-1048.
- [203] Raue S, Azzopardi L, Johnson C. #trapped!: social media search system requirements for emergency management professionals. SIGIR 2013: 1073-1076.
- [204] Yeniterzi R. Effective approaches to retrieving and using expertise in social media. SIGIR 2013: 1150.
- [205] Guy I, Zwerdling N, Ronen I, Carmel D, Uziel E. Social media recommendation based on people and tags. SIGIR 2010: 194-201.
- [206] Wang J, Li Q, Chen Y. User comments for news recommendation in social media. SIGIR 2010: 881-882.
- [207] Pavlidis Y, Mathihalli M, Chakravarty I, et al. . Anatomy of a gift recommendation engine powered by social media. SIGMOD 2012: 757-764.
- [208] Zhuang J, Mei T, Hoi S, Hua X, Li S. Modeling social strength in social media community via kernel-based learning. ACM Multimedia 2011: 113-122.
- [209] Qi G, Aggarwal C, Huang T. Community Detection with Edge Content in Social Media Networks. ICDE 2012: 534-545.
- [210] Popescu A, Shabou A. Towards precise POI localization with social media. ACM Multimedia 2013: 573-576.
- [211] Zafarani R, Liu H. Connecting users across social media sites: a behavioral-modeling approach. KDD 2013: 41-49.
- [212] Zhou A, Qian W, Ma H. Social media data analysis for revealing collective behaviors. KDD 2012: 1402.
- [213] Xu Z, Zhang Y, Wu Y, Yang Q. Modeling user posting behavior on social media. SIGIR 2012:

545-554.

- [214] Yin H, Cui B, Chen L, Hu Z, Huang Z. A temporal context-aware model for user behavior modeling in social media systems. SIGMOD 2014: 1543-1554.
- [215] Bergsma S, Durme B. Using Conceptual Class Attributes to Characterize Social Media Users. ACL 2013: 710-720.
- [216] Lampos V, Preotiuc-Pietro D, Cohn T. A user-centric model of voting intention from Social Media. ACL 2013: 993-1003.
- [217] Wang Y, Agichtein E, Benzi M. TM-LDA: efficient online modeling of latent topic transitions in social media. KDD 2012: 123-131.
- [218] Tran T. Exploiting temporal topic models in social media retrieval. SIGIR 2012: 999.
- [219] Zhou X, Chen L. Event detection over twitter social media streams. VLDB J. (VLDB) 23(3): 381-400 (2014).
- [220] Yin H, Cui B, Lu H, Huang Y, Yao J. A unified model for stable and temporal topic detection from social media data. ICDE 2013: 661-672.
- [221] Balahur A, Tanev H. Detecting Event-Related Links and Sentiments from Social Media Texts[C]. ACL (Conference System Demonstrations) 2013: 25-30.
- [222] Takase S, Murakami A, Enoki M, Okazaki N, Inui K. Detecting Chronic Critics Based on Sentiment Polarity and User's Behavior in Social Media. ACL(Student Research Workshop) 2013: 110-116.
- [223] Benson E, Haghghi A, Barzilay R. Event Discovery in Social Media Feeds. ACL 2011: 389-398.
- [224] Volkova S, Wilson T, Yarowsky D. Exploring Sentiment in Social Media: Bootstrapping Subjectivity Clues from Multilingual Twitter Streams. ACL 2013: 505-510.
- [225] Yin W, Mei T, Chen C. Automatic generation of social media snippets for mobile browsing. ACM Multimedia 2013: 927-936.
- [226] Xie L, Natsev A, Kender J, Hill M, Smith J. Visual memes in social media: tracking real-world news in YouTube videos. ACM Multimedia 2011: 53-62.
- [227] Stajner T, Thomee B, Popescu A, Pennacchiotti M, Jaimes A. Automatic selection of social media responses to news. KDD 2013: 50-58.
- [228] Jin X, Wang C, Luo J, Yu X, Han J. LikeMiner: a system for mining the power of ‘like’ in social media networks. KDD 2011: 753-756.
- [229] Lee C, Croft W. Building a web test collection using social media. SIGIR 2013: 757-760.
- [230] Li J, Liu C, Islam M. Keyword-based correlated network computation over large social media. ICDE 2014: 268-279.
- [231] Alonso O, Khandelwal K. Kondenzer. Exploration and visualization of archived social media. ICDE 2014: 1202-1205.
- [232] Guo W, Li H, Ji H, Diab M. Linking Tweets to News: A Framework to Enrich Short Text Data in Social Media. ACL 2013: 239-249.
- [233] 杨博, 陈贺昌, 朱冠宇, 赵学华. 基于超链接多样性分析的新型网页排名算法[J]. 计算机学报, 2014, 37(4): 833-847.
- [234] 黄立威, 李德毅, 马于涛, 郑思仪, 张海粟, 付鹰. 一种基于元路径的异质信息网络链路预测模型[J]. 计算机学报, 2014, 37(4): 848-858.
- [235] 韩毅, 许进, 方滨兴, 周斌, 贾焰. 社交网络的结构支撑理论[J]. 计算机学报, 2014, 37(4):

905-914.

- [236] 曹玖新, 吴江林, 石伟, 刘波, 郑啸, 罗军舟. 新浪微博网信息传播分析与预测[J]. 计算机学报, 2014, 37(4): 779-790.
- [237] 郭静, 张鹏, 方滨兴, 周川, 曹亚男, 郭莉. 基于 LT 模型的个性化关键传播用户挖掘[J]. 计算机学报, 2014, 37(4): 809-818.
- [238] 周东浩, 韩文报. DiffRank: 一种新型社会网络信息传播检测算法[J]. 计算机学报, 2014, 37(4): 884-893.
- [239] 张伯雷, 钱柱中, 王钦辉, 陆桑璐. 面向目标市场的信息最大覆盖算法[J]. 计算机学报, 2014, 37(4): 894-904.
- [240] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752.
- [241] 赵之灌, 于海, 朱志良, 汪小帆. 基于网络社团结构的节点传播影响力分析[J]. 计算机学报, 2014, 37(4): 753-766.
- [242] 毛佳昕, 刘奕群, 张敏, 马少平. 基于用户行为的微博用户社会影响力分析[J]. 计算机学报, 2014, 37(4): 791-800.
- [243] 赵秋月, 左万利, 田中生, 王英. 一种基于改进 D-S 证据理论的信任关系强度评估方法研究[J]. 计算机学报, 2014, 37(4): 873-883.
- [244] 王玙, 高琳. 基于社交圈的在线社交网络朋友推荐算法[J]. 计算机学报, 2014, 37(4): 801-808.
- [245] 何鹏, 李兵, 杨习辉, 熊伟, 陈军. Roster: 一种开发者潜在同行推荐方法[J]. 计算机学报, 2014, 37(4): 859-872.
- [246] 苑卫国, 刘云, 程军军. 微博网络中用户特征量和增长率分布的研究[J]. 计算机学报, 2014, 37(4): 767-778.
- [247] 赵旭剑, 杨春明, 李波, 张晖, 金培权, 岳丽华, 戴文锴. 一种基于特征演变的新闻话题演化挖掘方法[J]. 计算机学报, 2014, 37(4): 819-832.

作者简介

于戈 博士, 东北大学信息科学与工程学院, 教授, 博士生导师, 中国计算机学会理事, 办公自动化专业委员会主任。主要研究方向: 分布式数据管理、Web 信息系统等。E-mail: yuge@mail.neu.edu.cn。



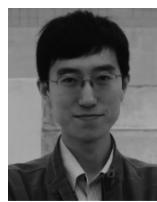
王大玲 博士, 东北大学信息科学与工程学院, 教授, 博士生导师, 中国计算机学会中文信息技术专业委员会委员。主要研究方向: 社会媒体挖掘、搜索与推荐等。E-mail: dlwang@mail.neu.edu.cn。



申德荣 博士，东北大学信息科学与工程学院，教授，博士生导师，中国计算机学会办公自动化专业委员会秘书长。主要研究方向：Web 数据管理、数据集成等。E-mail：shendr@mail.neu.edu.cn。



冯时 博士，东北大学信息科学与工程学院，讲师，中国计算机学会会员。主要研究方向：观点挖掘、网络舆情分析。E-mail：fengshi@ise.neu.edu.cn。



李瑞轩 博士，华中科技大学计算机科学与技术学院，教授，博士生导师，中国计算机学会办公自动化专业委员副主任。主要研究方向：大数据管理与分析，Web 数据管理，信息检索与数据挖掘。E-mail：rxli@hust.edu.cn。



李玉华 博士，华中科技大学计算机科学与技术学院，副教授，中国计算机学会会员。主要研究方向数据挖掘，社会网络分析，大数据挖掘分析，分布计算与云计算。E-mail：ideliyuhua@hust.edu.cn。



汤庸 博士，中山大学计算机学院，教授，博士生导师，中国计算机学会办公自动化专业委员会委员。主要研究方向：协同计算与社交网络，学术信息服务与大数据管理。E-mail：issty@mail.sysu.edu.cn。



邢春晓 博士，清华大学计算机科学与技术系，教授，博士生导师，中国计算机学会办公自动化专业委员副主任。主要研究方向：数据库，电子政务，数字图书馆等。E-mail：xingcx@tsinghua.edu.cn。



于 旭 (Jeffrey Xu Yu) 博士，香港中文大学系统工程与工程管理系，教授，博士生导师。主要研究方向：图挖掘、图查询处理、图模式匹配和关系数据库中的关键词搜索等。E-mail：yu@ se. cuhk. edu. hk。



姜安琦 硕士，2011 年加入新浪，现任新浪门户广告产品技术总监，主导推出了新浪 AdExchange、新浪龙渊、Adbox 等多款广告产品。加入新浪之前，在百度商务搜索部任职，主要负责百度展示广告相关产品研发，包括百度网盟、百度广告管家、百度品牌专区等。E-mail：anqi2@staff. sina. com. cn。



面向健康的感知与计算研究进展与趋势

CCF 微机专业委员会

周兴社¹ 王柱¹ 倪红波¹ 王天本¹ 林强²

¹西北工业大学计算机学院，西安

²西北民族大学数学与计算机科学学院，西安

摘要

近年来，随着信息技术不断进步，特别是感知技术与可穿戴设备的快速发展，为健康感知与计算相关研究的开展提供了技术基础。同时，现代社会各类慢性疾病蔓延，带来沉重的医疗保障负担，迫切需要革新当前的医疗与健康服务体系与技术，因此健康感知与计算受到世界各国政府越来越多的重视和投入。本文首先从健康感知设备与技术以及健康计算关键技术两个方面对健康感知与计算的发展现状进行介绍与分析，随后从多个角度对国内外研究进行了比较，给出了针对未来发展趋势的展望。

关键词：健康感知，健康计算，健康评估，健康促进

Abstract

With the constant progress of information technology, especially the recent surge of sensing technology and wearable devices, both the industry and academic communities have been paying close attention to the research of health-oriented sensing and computing. Meanwhile, the wide spread of chronic illnesses leads to a heavy economic burden to the modern society, and we must reform the current medical treatment and health care system and technology, which is attracting more and more attentions from countries all over the world. In this report, we first elaborate the development status and recent research results of health-oriented sensing and computing, and then prospects its future development trend.

Keywords: Health sensing, health computing, health evaluation, health promotion

1 引言

1.1 健康事关国计民生

健康不仅是人的基本权利，也是生活质量得以保证的前提和基础，体现着生命存在的良好状态。健康问题是我国社会当前所面临的重要而迫切的挑战之一。

一方面，随着现代化在我国的急速推进，随着社会节奏的加快，人们的生活和工作节奏也急剧加快。特别是以都市白领为代表的群体，由于生活节奏更快、工作压力更大、紧张度更高，长期处于高度精神紧张的状态之下，不能得到应有的调节，身心过度疲劳，从而导致焦虑不安、抑郁症、精神障碍等生理和心理健康问题。最近的一次调查表明，精神疾病已超过心血管病，跃居我国疾病患者的首位，约占 20%。根据世界卫生组织推算，到 2020 年我国精神疾病导致的医疗开销将上升至社会医疗总开销的 1/4。另据资料显示，我国 70% 的人处于亚健康状态，身心亚健康等现代文明病已成为多发病、常见病。

另一方面，我国已逐渐进入老龄化社会，老龄化问题成为影响国计民生和国家长治久安的重大战略性问题。根据《中国老龄事业发展报告（2013）》数据显示：2013 年全国 60 岁及以上老年人口已达 2.02 亿，占总人口的比重达 14.8%，是世界上唯一老年人口过亿的国家。根据《中国老龄事业发展“十二五”规划》预测，到 2015 年，全国 60 岁以上老年人口将增加到 2.21 亿，比重将增加到 16%。到 2030 年，全国 60 岁以上人口数将突破 4 亿，成为全球老龄化程度最高的国家。人口结构的老龄化必将对我国社会发展产生深远的影响，关注“银色群体”成为政府决策制定、社会经济发展及公共资源配置等方面优先考虑的全局性社会问题。相关健康统计结果表明，我国老年群体中患有一种或以上慢性疾病的人口比例明显偏高（比如心脏病、脑血管疾病及高血压），且呈现不断上升的趋势。由慢性疾病病因链可知，慢性疾病通常是由多种原因引起，同时具有病程长、早期症状不明显以及难以完全治愈等特点。一般而言，慢性疾病的发作往往与不良健康行为直接相关，及早发现并合理调控可在一定程度上控制或延缓其进一步发展。

随着现代社会人类生产与生活方式的深刻变革，特别是老龄化社会的到来，导致当前以疾病治疗为中心的医疗服务体系难以应对以慢性疾病为代表的现代文明病的快速蔓延，已经不能完全满足日益攀升的健康服务需求。因此，保障国民健康已成为重大的民生工程，需要全社会的共同关注与参与，例如，在社区和家居环境下、日常生活状态中实现健康监测、预警与辅助，通过探讨新的健康感知与计算理论与方法，使健康管理从被动治疗变为主动预防。

1.2 健康感知与计算是促进健康的重要手段

健康调节的前提是持续地感知和分析健康状态，并依据即时健康状况给予正确的干预和调节。外界的及时干预和调整需要建立在对人体健康状态正确感知与评估的基础之上。一般而言，健康评估不能单纯依靠个体的自身感知，也不能完全建立在对用户的问诊及化验体检基础上，而是需要日常生活环境下的持续性监测与分析。因此，为了促进和提升人体健康状态，需要研究持续性健康感知与计算，并实施有效的智能化健康辅助与促进服务。

随着信息技术不断进步，持续地进行人体健康状态感知、评估和预警已经成为可能。基于多途径的健康数据获取，进行持续多维度的分析并提取健康状态与各项健康数据之

间的内在关联关系，建立健康状态评估模型，实现有针对性的健康行为指导，对保持和提升人体健康水平、提高生活质量具有重要的理论意义和实际应用价值。

2 国际研究现状

2.1 健康感知国际研究现状

目前，国际上健康感知方面的研究工作主要集中在欧美。美国国家科学基金会（National Science Foundation, NSF）2012年设立的信息与智能系统项目群（Division of Information & Intelligent Systems, IIS）包含一批健康感知相关的项目^[1]，涉及可穿戴式生理参数感知及日常活动行为感知等内容。同时，健康也是欧盟第七研发框架计划（FP7）^[2]十大主题之一，其中非入侵式健康检查与监控是主要研究方向。下面从可穿戴和非可穿戴两个方面详细阐述国外在健康感知方面的研究进展。

2.1.1 基于可穿戴设备的健康感知技术

可穿戴设备是健康感知的重要手段之一，其具有两方面明显优势：1) 可长时间持续动态检测，获取丰富的健康数据；2) 可穿戴所带来的便捷性，使得健康参数采集可在日常生活环境下进行。因此，国外对可穿戴健康感知技术的研究可谓丰富多样，我们将可穿戴健康监测系统进一步依据外观和结构分为：基于体域网的感知设备、基于智能终端的感知设备以及基于智能织物的感知设备。

利用具有无线传输功能的生理传感器节点组成体域网（Body Area Network, BAN），检测人体生理参数是可穿戴健康感知的重要分支之一。阿拉巴马汉茨维尔大学研制的个人健康监测体域网系统^[3]使用集成有Zigbee无线通信模块和超低功耗微控制器的无线传感器开发平台Telos和用户可定制多传感器模块实现加速度、ECG及EMG信号的检测。每个感知节点均具有独立的数据感知、处理、传输功能。哈佛大学的CodeBlue系统^[4]同样以Telos平台为基础，用户定制传感器模块包含血氧饱和度、ECG、EMG和运动参数，该系统特别强调传感器节点及其与数据汇聚节点之间数据通信的可靠性。此类设备硬件平台通用性较好，具有较强的感知与计算能力。由于使用独立的感知模块，用户可按需定制传感器，但其要求低功耗和低辐射，而且器件的工艺较为复杂。

以智能手机为健康感知与数据汇聚平台的可穿戴设备越来越广泛的用于健康感知。微软的HealthGear系统^[5]的感知模块可同时监测用户血氧饱和度及脉搏。感知模块通过蓝牙将传感器数据传输至智能手机，通过具体的数据分析，可自动检测睡眠过程中呼吸暂停事件。哈佛-麻省理工卫生科学与技术部（Harvard-MIT Division of Health Sciences and Technology）和剑桥大学联合研制的健康检测系统Heartphones^[6]将心跳检测仪嵌入耳机中，实现心跳频率的检测，并以智能手机显示检测结果。三星Galaxy S5智能手机可通过

内置的心率传感器实现心率数据的采集，在经过健康服务软件的简单分析和处理后，将心率数据呈现给用户。此类设备一般具备可穿戴性较强，体积小巧，但可检测的健康数据种类较少。

在智能织物方面，哈佛大学研制的智能织物健康监护系统^[7]将光电容积扫描血压计和多导 ECG 电极嵌入到布料当中，实现血压和心电信号的实时检测。欧盟可穿戴健康监护系统项目（TheWearable Health Care System Project），为欧盟第五研发框架计划中的项目之一，通过在纺织布料中嵌入特殊材料电极和压电材料制成的传感器，实现 ECG、EMG、呼吸率、体温、血压、血氧等参数的持续检测，并可基于 GPRS 或蓝牙模块将预处理后的数据发送至健康服务器。佐治亚理工大学研制的通用可穿戴生理参数监控设备^[8]可兼容多种类型的生物传感器，实现多种生理参数的实时监控。印度国防研究与发展组织工程与电子医学实验室研制的远程生理监控智能背心^[9]，可同时检测穿戴者 ECG、PPG、心率、血压、皮电等生理参数，最终利用采集到的多项生理参数对身体健康程度进行评估。该设备的特点是使用硬件实现 ECG 信号的高通、低通、阶梯滤波，因而无需事先设定噪声基准及误差范围，且使用 ECG 数据辅助矫正血压数据。此类设备为可穿戴感知的高级形态，具有穿着舒适、不影响日常活动等优点，但工艺复杂，对制作材料有严格要求。

此外，国外产业界也研制出众多较为先进的健康感知产品。以可穿戴性最佳且目前最流行的智能手环为例，耐克公司研制的 Nike + Fuelband 智能手环可同时检测日常运动参数以及心率和脉搏等生理参数，并支持睡眠检测。监测到的数据可通过蓝牙 4.0 传输至 Android 智能终端进行查看。Fitbit 公司研发的智能手环可记录每天的行走路程、燃烧消耗、活跃时间、睡眠时长以及睡眠质量等并将数据同步至 iOS 或 Android 智能终端。另外，该手环一次充电可连续工作 5~7 天，避免频繁充电带来的不便。爱普生（Epson）公司 Pulsense 系列智能手环通过检测红细胞反射的光通量的变化来准确识别心跳。同时，该手环可通过其他传感器追踪用户睡眠情况。

虽然可穿戴设备在健康检测方面具有明显的优势，但是纵观已有可穿戴式健康监测设备，软硬件开发平台种类繁多，缺乏统一通用的开发平台；其次，在可穿戴设备与佩戴者之间，绝大部分研究限于特定参数感知，缺乏方便的人机交互支撑；最后，除智能织物类感知设备的穿戴舒适感接近真实衣物配饰之外，其他类型可穿健康感知设备的舒适性还有待提高。

2.1.2 基于非可穿戴设备的健康感知技术

以非可穿戴设备为载体的健康感知技术旨在研发能够方便部署于家庭生活环境之中的健康感知系统，典型的代表包括智能床垫、智能座椅等。相比而言，以非可穿戴设备为载体的健康感知系统避免了电池续航能力、计算能力、数据存储能力等限制。本文按照感知原理的不同将非可穿戴式感知技术分为：基于压电技术的健康数据感知、基于音频与图像技术的健康数据感知以及基于光纤技术的健康数据感知。

压电传感器是获取人体健康数据的常用感知设备之一，通过在床垫或沙发等生活设施上布设压力传感器捕获人体微动产生的压力，可实现呼吸及心跳等健康数据的感知。

如亚琛工业大学^[12]研究基于压力传感器检测 BCG 信号，并从 BCG 信号中提取出心率数据；该系统使用 4 个置于床垫适当位置的应变仪捕获心肌运动产生的压力波信号，原始信号进一步处理后可提取出心率数据。东京大学^[13]通过在枕头中布设压力传感器矩阵监测用户在睡眠过程中头部对枕头的压力的变化，并提出一个根据压力变化检测呼吸事件的模型，实现睡眠过程中呼吸事件的监测。为增强压力传感器的灵敏度，日本会津大学^[14]通过将压力传感器与流体压敏设施（如空气或液体床垫）结合可实现呼吸及心率等健康数据的准确感知。由于压力传感器的电磁敏感特性，外界电磁干扰容易引起感知数据不准确甚至错误。

在基于音频与图像技术的健康数据感知方面，帕多瓦大学^[15]研究了基于光学传感器的非侵入式感知设备，采集人体皮肤及皮下组织的生物-物理特征数据，应用于糖尿病患者血糖含量的评估。全北大学^[16]、普纳大学^[17]研究了利用麦克风采集用户在睡眠过程中的呼吸声音，识别鼾症，并通过内置于枕头中的充气气囊迫使用户改变睡姿，防止睡眠呼吸暂停。视频、音频健康数据感知技术对应用环境具有较高的要求（如光照条件、噪声水平等），数据后期处理过程较为复杂。同时，基于图像技术的检测技术一般需要被检测者主动参与感知过程，一定程度上限制了其适用性。

光纤具有无辐射、免电磁干扰等优良特性，因此特别适合于日常生活环境下的长期健康数据感知。从技术角度讲，光纤健康数据感知设备分为光强传感器和光波传感器两类。其中，光强传感器利用光纤光强对压力的敏感特性原理实现人体微动（如呼吸、心跳和体位移动）的检测，新加坡科技研究局通讯研究院^[18]研究利用布设在床垫和枕头上的光强光纤传感器分别实现呼吸与身体移动及心跳事件的检测；光波传感器基于光纤光波对外界作用力的敏感特性实现包括人体体温、呼吸、心跳和身体位移的检测，如新加坡科技研究局通讯研究院^[19]研究利用布设在床垫上的光纤布拉格光栅传感器串联阵列实现呼吸和心率数据检测；马里博尔大学^[20]研究基于光学干涉仪的呼吸和心跳事件检测，呼吸和心跳引起的人体微动被干涉仪捕获后转换为光纤波长的变化信息，从而实现呼吸率和心率健康数据的非干扰感知。光强光纤传感器对温度不敏感，因此不能用于感知人体体温数据；光纤光栅传感器可感知外场温度和应力变化，但合成感知数据的不同分量的分离仍是该领域面临的挑战。

非可穿戴式健康检测设备一般布置于日常使用的物件或局部环境当中，具有不受供电、计算能力、存储能力限制等优势。但是由于设备本身物理特性的限制，不具备移动性或移动性较差，受限于使用者具体行为，此类设备只适合在特定时间间隔内的健康感知与监测。

2.2 健康计算国际研究现状

健康计算旨在通过对健康数据特别是健康数据流进行长时间、持续性的分析、建模与评估，发现体征参数异常、识别并评估健康状态以及预测健康发展趋势，为促进人体健康提供依据。目前，国际上健康计算领域的研究主要围绕健康数据分析、健康状态建

模与评估以及健康促进等方面展开。

2.2.1 健康数据分析方法

一般而言，随着年龄的增加人的生理与认知能力呈现缓慢而持续的下降，慢性病人群尤其如此。因此，通过多种感知方式所获取的健康数据以数据流为主要形态，具有鲜明的复杂、多源、异构、多维、持续且共同进化等特点。以电子健康记录（Electronic Health Records, EHR）为例，其既包含结构化数据，也包含半结构化和非结构化数据。目前健康数据分析的研究热点主要包括面向数据流的持续性实时分析与多维度关联分析。

健康数据分析的主要目的是检测异常事件的发生并准确预警。基于传统的非持续性健康数据，国外学者开展了大量的研究，例如文献^[21]通过分析 ECG 信号的动态性特征，检测心律失常事件的发生；文献^[22]通过综合分析 EEG、EOG、EMG、ECG 等数据，检测阻塞性睡眠呼吸暂停事件的发生。目前，这一研究领域的难点主要体现在如何围绕持续性多数据流设计高效的分析与挖掘方法，也已取得了一些阶段成果。Zhu 和 Shasha^[23]提出了针对时间序列数据流进行统计性计算的技术，通过一个随机选定的滑动窗口，研究数据流快速检测方法。Guralnik 和 Srivastava^[24]提出了一种面向时间序列数据流的事件检测方法，对多个共同演化的时间序列采用多变量线性回归的方法，虽然能够处理半无限序列，但是对于数据流的增量式累积显得无能为力。

在健康数据多维度关联分析方面，目前流行的方法主要有两种类型：统计方法和数据挖掘方法（监督、半监督、非监督）。经典的统计方法中，常用的有 Pearson 检测和 Fisher 检测，但其对数据的标准化和一致性要求较高，健康数据的上述特点导致其并不适用于多维度关联分析。基于数据挖掘方法的研究中，Batal 等^[25]提出了用于发现健康数据中时间模式的方法，其核心是基于时间抽象对健康数据进行表示，具体采用 Apriori 算法实现。文献^[26]基于重构分析技术（Reconstructability Analysis, RA），分析研究不同维度健康数据之间的关联关系，发现所蕴含的强相关子集并基于简化模型进行表示。Apriori 等经典数据挖掘方法虽然能够有效处理分段式定性数据，但是不能针对连续性数值数据进行关联分析，对持续性的累积效应也不能很好地测度，因而不能适用共同演化复杂数据流的相关性分析。

综上可知，目前对于数据流的分析大多面向单数据流，且已经提出部分有效算法，但是针对多数据流的分析有待进一步研究。结合健康数据流的特点，需要重点研究适合多维度、持续性、共同演化数据流的异常检测与关联分析方法，克服通用数据分析方法应用于健康数据分析所存在的缺点，为人体健康评估提供定量化科学依据。

2.2.2 健康状态评估模型

健康状态建模的目的是通过分析健康数据发现数据特征与特定健康问题之间的关联模式，从而准确刻画、评估和预测人体的健康水平，其核心基础是对健康问题的形式化建模。目前，国际上健康状态建模的相关研究，从建模方法的角度，可分为基于经验数据的统计模型与基于机器学习的概率模型；从所针对健康问题的角度，可分为面向具体疾病的健康模型与面向人体整体健康状态的模型。

在健康状态评估模型的构建方法方面，部分学者研究基于历史经验数据的统计模型，

即依赖若干生理参数，构建健康水平评分体系，如针对糖尿病的 PreDxH Diabetes Risk Score 与 MetS 体系^[27]等，针对心脏病的 Framingham Risk Score 与 SCORE 体系^[28,29]等。更多的学者研究基于机器学习理论与方法的概率模型，如 Tresp 等^[30]研究基于神经网络建模糖尿病患者的血糖水平，Wei 等^[31]则基于支撑向量机对 I-型糖尿病进行建模与评估；Palaniappan 等^[32]及 Dangare 等^[33]人分别研究基于朴素贝叶斯、神经网络、决策树三种方法构建心脏病评估模型，并对不同建模方法的性能进行了比较；Wu 等^[34]则基于逻辑回归分析、支撑向量机以及 Boosting 三种方法构建心力衰竭评估模型，亦对不同建模方法的性能进行了分析与比较。

在面向特定健康问题的评估模型方面，国外学者主要研究具体的疾病，其中研究较多的有糖尿病、心脏病等，例如 Noble 等^[35]以及 Abbasi 等^[36]分别从不同的角度系统地分析与比较了 145 个与 25 个针对 2-型糖尿病的评估模型，指出了目前研究的不足与可能的研究方向；Siontis 等^[37]则系统分析了 8 个针对心脏病的评估模型。此外，部分国外学者从人体系统或子系统的角度，研究健康状态的评估模型。例如，文献^[38]研究基于心率变异性（Heart Rate Variability, HRV）分析构建针对中枢神经系统的健康状态评估模型，其基础是心率变异性降低与中枢和周围神经系统疾病的关联关系。SCAI 组织^[39]开展了基于 ECG 信号的人体复杂性与变异性建模研究，成为预测心血管系统疾病及其并发症的主要方式之一。

总体而言，国外现有研究一般针对具体的慢性病构建模型，故而只能对人体某个器官或子系统的状态进行评估，而不能从宏观的角度评估人体系统的整体健康水平，无法为健康促进提供全方位、多粒度的信息。同时，人体系统理论认为健康的理想境界是人体各子系统正常运行且整体系统处于均衡状态，特别是现代社会中亚健康人群规模不断增大，如何通过健康评估和促进使身体保持均衡状态就显得尤为重要。因此，需要从人体系统健康的角度构建层次化模型，实现在宏观与微观两个粒度全方位评估人体健康状态。

2.2.3 健康促进理论与方法

健康促进（Health Promotion）一词始见于 20 世纪 20 年代公共卫生领域的文献^[40]，其内涵主要包括个人行为改变和政府行为（社会环境）改变两个方面，重视发挥个人、家庭和社会的健康潜能。人类健康实践证明，健康促进是人类应对现代社会生活方式病和慢性病蔓延挑战，满足社会老龄化趋势需求，解决“看病难、看病贵”，以及提升人力资本和国民素质的科学有效途径。

在准确评估健康状态的基础上，部分学者进一步开展了健康促进理论与方法方面的研究，如 Pender 等^[41]提出的 HPM 模型，通过整合护理和行为医学领域的概念性架构，归纳出影响健康行为的因素，主张健康促进取决于认知 - 知觉因素和修正因素；Dahlgren 等^[42]提出社会经济模型，主要从社会、经济、文化、环境等角度分析影响人体健康的因素；Beattie 等^[43]从健康干预的模式（命令式、协商式）与对象（个体、群体）两个方面定义了健康促进的四种表现形式；Green 等^[44]则从健康促进规划的角度研究健康促进，将健康促进划分为 9 个阶段。基于健康促进相关理论模型，学界从多方面开展了健康促

进方法的研究，一是探索健康促进的不同途径，二是研究如何提升健康促进的有效性。

在健康促进途径方面，目前国外的相关研究主要包括基于行为改变的健康促进与基于社会网络的健康促进。在基于行为改变的健康促进研究中，Consolvo^[45]等人研发了UbiFit系统，基于穿戴式感知、实时活动推理等技术鼓励用户在日常生活中进行有规律的体育活动，以促进身体健康；Albaina^[46]等人研发了Flowie系统，通过“虚拟教练”督促用户更多的走路；Grimes^[47]等人研发了OrderUP！系统，通过游戏改变人的饮食习惯，促进身体健康。在基于社会网络的健康促进研究中，Munson^[48]等人通过在社交网络中嵌入健康干预应用，促进人体健康；Oliveira^[49]等人研发了MoviPill系统，通过社交网络游戏督促用户按时服药；Newman^[50]等人则系统地分析了基于社交网络促进健康所面临的挑战与机遇。

在健康促进有效性方面，国外学者主要从个性化与情境敏感（Context- Aware）两个方面开展研究。针对健康促进的个性化，Jones^[51]等研究面向患者的个性化信息服务系统，基于用户的个性需求对信息进行裁剪，为不同用户自适应呈现其感兴趣的内容。针对情境敏感的健康促进，Vurgun^[52]等以及 Hayes^[53]等研究了情境敏感的用药提醒系统，基于动态贝叶斯推理当前情境下（如时间、地点、用户活动等）是否适合向用户进行提醒，以有效促进其健康。

3 国内研究进展

3.1 国内健康感知与计算主要研究项目

近年来，在国家重点基础研究发展计划（973计划）与国家自然科学基金的支持下，国内启动了数个健康感知与计算领域的重大研究课题。代表性科研项目包括：

（1）基于生物、心理多模态信息的潜在抑郁风险预警理论与生物传感关键技术研究

科技部国家重点基础研究发展计划项目，2014年启动，2018年截止，依托兰州大学，胡斌教授为首席科学家。

近年来，随着社会对精神疾病关注程度的不断提升，尤其是普通人群中抑郁症患者的快速增加以及传统监测治疗方法局限性的日益突出，迫切需要建立新型的抑郁类精神疾病辅助诊疗系统。与此同时，普适计算、生物信号反馈等信息技术的迅猛发展，使得研制此类个性化、智能化、普适性的系统成为可能。该项目主要通过获取多模态健康数据，进行心理干预与诊疗等方面的研究。一方面，针对随机性和背景噪声较强的生理信号，开展特征提取、特征识别、数据压缩和自动分类等方面的研究，为多模态特征融合、建模以及生理计算奠定基础。另一方面，通过对生理信号、医学影像以及个人基本信息等多模信息的表达、组织与建模，研究适用于不同人群（如抑郁症患者）的模型，以便

准确、客观地监控异常情感与心理状态变化，并及时合理地进行心理干预与治疗。

(2) 面向老年人健康的非干预式感知与持续计算研究

自然科学基金重点项目，2014 年启动，2018 年截止，依托西北工业大学，周兴社教授为项目负责人。

该项目面向我国老年人健康特点与需求，基于人体系统论与信息论，以面向老年人健康的感知与计算方法及其关键技术为主题，重点研究非干预式健康状态感知、持续性健康数据多维度关联分析、人体健康评估模型及其促进方法。项目注重解决感知与分析方法以及评估模型的准确性、高效性以及适应性等基本问题；同时，通过构建面向老年人健康数据的感知与分析原型平台，实施典型实例验证，评测和提升研究成果的有效性，为服务我国老年人健康生活提供方法与技术支撑。

(3) 面向人类健康的体外诊察信息感知与计算方法研究

自然科学基金重点项目，2014 年启动，2018 年截止，依托香港理工大学深圳研究院，张大鹏教授为项目负责人。

该项目针对目前可体外诊察感知的状态单一有限以及医学诊断现代化研究中面临的数字化、标准化、可重复性和可扩展性等瓶颈问题，从具有广度的全方位体外诊察感知入手，旨在实现面向人类健康的多层次信息融合计算，建立面向人类健康的体外诊察信息分析应用平台，从而推动我国医学诊断现代化领域源头性创新。

除上述项目之外，科技部国家高技术研究发展计划对数字化医疗工程技术领域给予了重点支持，旨在开发基于体域网的个人健康信息智能采集技术及系统；国家科技支撑计划则重点支持了健康云平台、健康服务业等相关领域。此外，国家自然科学基金委在 2014 年重点资助基于可穿戴计算的情感交互理论与方法等面向心理健康的研究项目。

3.2 健康感知与计算国内研究动态与进展

近年来，在各类科研项目支持下，我国的研究团队结合各自的研究基础与应用方向，在健康感知与计算领域取得了一些进展。

3.2.1 健康数据感知与分析

保障人体健康的前提是获取健康状态信息，从 IT 技术的角度出发，既不能单纯依靠个体的自身感知，也不能完全基于对用户的医疗问诊和化验体检，而是应该强调日常生活环境下的自然感知与监测。同时，基于不同方式获取的健康数据，往往具有各自的特性，需要采用不同的方法进行处理和分析。

结合各自所承担的国家重点课题，香港理工大学围绕视觉、听觉、嗅觉和触觉等四个方面开展体外诊察研究，旨在实现基于多层次信息融合的健康感知与计算。例如，通过感知人体呼吸的化学成分，研究相关特征与糖尿病的关联关系^[54]；基于多核学习提取并融合不同脉象特征，研究相关特征与疾病的关联关系^[55]；基于舌苔和人脸特征，构建相关病症的自动分析系统^[56]。兰州大学针对随机性和背景噪声较强的生理信号开展了感知与分析方面的工作，包括脑电、心电、肌电、眼电、神经电等生物电信号，以及呼吸、

脉搏、血压、血流、温度等生理量。西北工业大学利用光纤传感器对外场应力的敏感机理及其电磁不敏感和无辐射等优良特性，研究在自然睡眠状态下持续感知体温、呼吸、心率和身体位移等基本健康数据的机理和方法。

此外，在基于射频技术的健康数据感知与分析方面，第四军医大学^[65]、北京航空航天大学、西安电子科技大学^[66,67]等开展了呼吸与心率数据检测方面的研究。在基于图像与视频技术的健康数据感知与分析方面，厦门大学^[61]研究通过感知与分析用户脸部、眼部与舌苔的图像和视频信息，评测健康状态；类似的研究还包括哈尔滨工业大学^[68]围绕齿痕舌图像信息的感知与分析以及西南交通大学^[69]面向心电、体温以及呼吸信息的感知与分析。在基于各类新型传感设备的健康数据感知与分析方面，中科院计算所基于传统医学原理，研发了脉搏感知系统，提出了相应的信号处理和特征提取方法，能够准确感知脉搏数据并评估人体的健康状况^[57]；清华大学通过感知血糖、血压、血脂等时序健康数据，研究了基于时间序列距离度量的异常事件检测方法，为糖尿病等慢性疾病的管理提供辅助^[58]；空军航空医学研究所以检测呼吸睡眠事件为目标，研制了微动敏感床垫睡眠检测系统，实现了无电极条件下高分辨率生理信号的获取，对用户的睡眠影响很小，且能准确评价其睡眠状态^[70]。

3.2.2 健康状态建模与评估

近年来，在健康状态建模与评估方面，国内研究机构已开展了一些创新性的研究工作。例如，兰州大学针对精神类疾病和心理健康问题，通过表达、组织与建模生理信号、医学影像、个人基本信息等多模态数据，并结合数理统计、模糊数学及语义表达等理论和技术，研究适用于不同人群的心理健康状态评估模型。此外，部分国内学者通过对国外学者提出的健康状态评估模型进行修正和完善，提出适合我国人群特征的健康模型。例如，浙江大学^[59]综合影响心血管健康的各种因素，借鉴成熟的风险评估体系Framingham Risk Score与SCORE，并结合专家知识构建适合中国人特征的心血管健康状态评估模型；上海交通大学^[60]则针对现有模型与方法的不足，提出了基于规则引擎的健康状态评估系统。

另一方面，受传统医学理论影响，国内部分学者倾向于从系统的角度建模与评估人体的整体健康水平，例如中科院计算所、厦门大学等分别从不同角度研究人体健康状态（健康、亚健康、疾病等）的评估模型^[57,61]。航空医学研究所^[62]研究了睡眠周期及其结构特征，并从中医的角度对不同特征与健康问题之间的关联关系进行解读，构建相应的人体健康评估模型。

3.2.3 健康促进

目前，国内学者对于健康促进定义的基本共识是“以教育、组织、政策和经济等手段，干预对健康有害的生活方式、行为和环境，以促进人体健康”。具体而言，健康促进旨在改变不健康的行为，改进健康预防性服务以及创造良好的社会与自然环境，其内容包括政府立法，解决有害的生产、生活环境，支持和促进个人、家庭和社会共同承担卫生保健责任；增加与改善健康预防性服务设施，投入资源以促进国民健康；倡导文明、健康、科学的生活方式，提高国民的自我保健意识和技能。浙江大学^[63]从公共健康卫生

的角度出发，开展了健康促进模式的研究，提出了健康促进适宜性选择模型。兰州大学针对精神疾病类人群日益增加的现实，研究面向不同心理问题人群（如抑郁症患者、心理高压患者和轻度认知障碍患者等）的自适应干预机制和治疗方法。西北工业大学^[64]针对我国社会快速老龄化的现状，研究了情境敏感的智能化辅助系统。例如，通过实时检测用户的位置、活动及可用设备等情境信息，在合适的时间以适合的多媒体形态提供用药提醒服务，避免药物误服和重复用药等问题的发生，从而保障老年人的身体健康。

4 国内外研究进展比较

目前，国内外在健康感知与计算领域的研究所存在的不同主要体现在以下几个方面。

(1) 健康感知与计算的系统论

目前，国外健康感知与计算领域的总体研究水平虽然领先国内，但是相关研究一般围绕具体的健康问题展开，或者关注健康数据的获取与分析，或者关注健康状态的建模与评估，尚未形成整体性、系统化的理论体系。中华民族源远流长，在数千年历史长河中，先贤们积累了许多被证明是卓有成效的人体健康保障方法。例如，《黄帝内经》给出了人体的系统观与均衡论。从身体调节和健康保持的角度出发，我国著名健康学专家俞梦孙院士结合中医理论，提出了如图1所示的人体健康调节模型，认为健康状态是一种机体平衡状态，外界的及时干预和调节对保持机体的平衡具有重要作用。该模型以身心健康状态的感知（Sensing）、辨识（Identifying）与调理（Regulating）为主要内容，其中感知旨在获取与人体整体健康状态有关的信息，为健康状态辨识提供数据基础；辨识旨在发现健康问题的方向、层次和程度，从而决定具体的调理方式；调理的目的则在于通过心理、饮食、锻炼以及理疗等各种方式、方法，促进人体平衡状态的恢复和保持。

(2) 健康数据的感知方式与深度

健康数据感知的相关研究，国内外的差别主要体现在感知方式与深度的不同。基于智能织物的健康感知方面，国内研究起步较晚，与国外的差距较大。目前国外虽尚无商业化的智能织物产品，但原型系统众多，技术丰富且实现技术先进；国内由于受材料、工艺等方面的限制，智能织物研究目前处于初步探索阶段。基于体域网的健康感知方面，国内外研究水平相当。美国哈佛大学等知名高校起步较早，具有相对较好的技术储备；国内由科技部所资助的一批基于体域网的个人健康数据采集相关项目正处于在研之中。另外，清华大学、中科院计算所、解放军总医院、吉林大学等已在该领域取得阶段性研究成果。基于智能手机的健康感知与服务方面，国外仍处于领先地位，三星、苹果等智

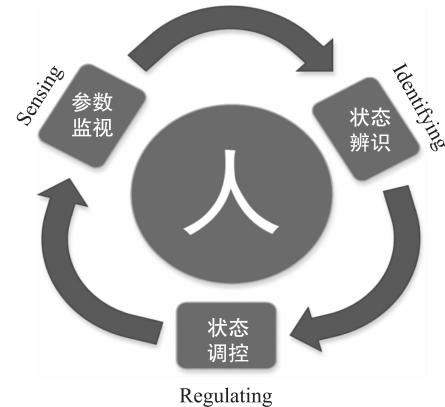


图1 基于系统论的健康调节模型

能手机巨头均拥有各自的商业化智能手机健康应用。目前多数此类健康应用一般采用外置的第三方传感器实现生理数据的感知，而三星已将生理传感器嵌入到其最新研发的智能手机中。国内方面，目前并未有智能手机生产厂商推出健康检测相关的产品或应用。此外，在基于压电、光纤技术的健康数据感知方面，国内外研究水平相当。国外有新加坡科技研究局通讯研究院、东京大学等为首的科研单位或高校，国内则有航空医学研究所、西北工业大学等开展相关研究，且均已推出原型系统或阶段性成果。

(3) 多源健康数据的分析方法

在健康数据特别是健康数据流的分析方面，国外相关研究起步较早，通过借鉴数据流挖掘领域的最新成果，已形成较完整的方法和技术体系，目前研究的重点是如何克服通用数据分析方法应用于健康数据分析所带来的缺点。国内在健康数据分析方面的研究虽然起步较晚，但在相关课题的支持下取得了较快的发展，目前部分研究已经达到国际先进水平，而且具有显著的中国特色。例如，在睡眠结构分析方面，国外研究多数基于多导睡眠图获取和分析用户的脑电、眼电及肌电等数据，此种方法虽然精度和可靠性较高，但是一般用在医院中，并不适合日常家庭环境；航空医学研究所创新性地提出基于较易获得的心动、呼吸、体动等基本生理参数研究睡眠结构，并采用模糊推理进行多睡眠数据的融合计算，实现了自然状态下无电极、无约束的睡眠监测与分析。香港理工大学结合“望闻问切”的传统医学理念，从视觉、听觉、嗅觉和触觉四个角度开展了全方位的体外诊察感知与计算研究，提出了面向人类健康的多层次信息分析方法与融合机制（包括感知融合、特征融合、匹配融合及决策融合），有助于推动我国医学诊断现代化领域的原始创新。

(4) 健康状态评估模型的构建角度与解读方式

由于受现代医学专科化、精细化趋势的影响，多数国外学者侧重构建面向特定健康问题或疾病的评估模型，其主要目的是为了预测具体健康问题的发生概率，并提供有针对性的健康服务。一般而言，所服务的对象即为模型所评估的对象本身。换言之，国外学者多数遵从微观的解读方式，将人体不同子系统或器官看做孤立的存在。国内的学者，特别是以俞梦孙院士为代表的一些科技工作者，由于受到中国传统医学理论中人体系统观与均衡论的影响，更加倾向于从系统的、综合的、整体的角度构建健康评估模型，所评估和预测的对象是人体的整体健康水平而不是某一个具体的子系统或者器官。对于健康问题的理解，则基于复杂系统理论，特别是自组织原理，从宏观的角度进行解读，认为具体健康问题的出现是人体整体失调的局部体现。此时，健康服务将提供给引发健康问题的根源，而不是问题的表象，因此更可能从根本上提升人体的健康水平。此类基于人体系统观理论而构建的健康评估模型尤其适合以各类慢性病为代表的现代文明病。

综合上述国内外研究进展可知，在面向健康的感知与计算这一新型研究领域，我国虽然起步较晚，但是在相关课题的支持下取得了较快的发展，目前部分研究已经达到国际先进水平，特别是在健康感知与计算的系统论、健康状态评估模型的构建角度与解读方式等方面形成了显著的中国特色，有可能为克服各类慢性疾病、服务老龄化社会、提升人类健康水平与生活质量做出突破性贡献。

5 发展趋势与展望

5.1 可穿戴设备与健康感知相结合

可穿戴设备的持续工作特性及其与生俱来的便携性为健康感知带来了新的发展契机和挑战。可穿戴设备和健康感知相结合的首要优势是能够实现长时间的动态监测，提供丰富的健康感知数据，有利于实现健康状态的客观评价，并及时实施健康促进。传统的一次性检测形成的部分人体生理指标很难得出准确客观的结论，只有通过长时间的持续检测才能得到相对可靠的测量结果。例如，在早期心脏病监测中，一次心电图难以捕捉到有效的诊断依据；而症状最明显的时刻往往是心电图采集的最佳时机，但是实际中由于检测的不连续性，此类时机往往被错过。基于可穿戴设备的动态心电图监测可持续的记录受测者的心脏活动状况，包括休息、活动、进餐、工作、学习和睡眠等不同活动下的心电图数据，从而发现一次性常规心电图不易发现的心律失常、心肌缺血等健康问题。随着我国人口老龄化趋势的不断加剧，可穿戴设备和健康感知相结合的另一个优势是可以避免慢性病患者频繁就医，节省开销。目前大多数慢性病患者需要定期就医，检查相关生理指标，明确病理发展趋势。然而，频繁的就医必然带来人力和财力方面的负担，如果能够将可穿戴健康感知设备应用于慢性病检测则可实现远程生理指数采集及诊断，从而减少就医次数，节省医疗开销。可穿戴计算应用于健康感知虽然具有诸多优势，但仍然需要克服下述挑战。

其一是传感器感知能力的准确性和可靠性。由于可穿戴设备对传感器尺寸及结构的限制，相比于医学领域认可的测量标准，目前嵌入在可穿戴设备的生理传感器在准确性和可靠性方面仍具有较大差距。例如，目前医用 ECG 监护仪多为 12 导电极，而使用在便携式设备或可穿戴设备上的 ECG 检测仪多数采用 3 导电极，检测精度仅能满足较为粗略的健康评估。

其二是设备的穿着舒适性。目前，智能织物之外的其他类型可穿戴感知设备的穿着舒适性均较差，会对用户的日常行为造成一定程度的干扰。对于用户而言，如果可穿戴健康感知设备仅停留在“可以穿戴”而非“适合穿戴”的层面，则会阻碍可穿戴健康感知设备的发展和普及。

5.2 大数据与健康状态评估相结合

健康数据是典型的大数据，近年来大数据研究的兴起为开展健康计算特别是健康状态评估模型研究提供了前所未有的机遇，二者的结合已经成为了健康计算研究的必然趋

势。只有合理分析和挖掘健康大数据，才能构造具有普遍适用性的健康状态评估模型，并进一步以较低的成本为不同健康问题和不同特征人群提供系统化与个性化相辅相成的健康评估模型和健康促进服务。然而，区别于其他大数据，健康大数据的一些特有性质，给二者的结合带来诸多挑战。

其一，健康大数据具有多源、异构、动态等特点，而且是持续增长的大数据。因此，高效可用的健康评估模型不但需要具有良好的可扩展性，以表达、组织和建模不同类型的数据；同时，还需要具有良好的自适应动态进化能力，随着人体健康状态的动态变化进行相应的挑战和演化，从而实现持续而准确的评估。

其二，健康大数据是关系复杂且富含语义的多维数据。因为数据来源的多样性，不同维度数据之间的关系可能非常复杂。同时，健康大数据蕴含了丰富的语义信息，准确发现相关语义是构建健康评估模型的基础。另外，不同的健康服务可能需要从不同的视角分析和解读数据。如何研发适合健康大数据的数据挖掘和分析工具，揭示健康数据所蕴含的语义信息及其与健康状态之间的关联关系，并对相关结果进行深度而合理的解读，是构建健康评估模型时需要深入探讨的问题。

其三，健康大数据涉及隐私问题。随着健康大数据时代的来临，普通大众获取健康信息的渠道和内容发生巨大变化，健康数据的质量和安全问题也日益凸显。虽然大数据有利于构建更加有效的健康评估模型，但是如何保证建模过程中用户隐私不被泄露或恶意利用，是大数据与健康状态评估相结合所需面临的又一挑战。

5.3 健康服务的 Online-to-Offline (O2O) 模式

健康感知与计算研究最终需要以健康服务的形式提供给用户。因此，在服务模式方面，需要进一步研究适合中国国情的线上与线下相结合的服务模式，整合分散的需求与供给，形成一条完善的健康服务供应链，将服务方和服务对象进行有效的链接，进而形成一种长效发展的新型健康服务模式，以满足不同人群对健康服务的需求。一方面通过线上实现服务资源整合，提供多样化健康信息服务，另一方面依托线下实施具体的健康服务，形成服务对接优势，从而构建线上线下一体化、多渠道资源整合的健康服务管理平台。目前，O2O 健康服务模式的进一步发展还需解决以下关键问题。

其一，需要进一步整合健康服务云平台与智能感知终端，将健康感知、计算与服务紧密结合。一方面，依据不同群体，部署或采用与其适应的健康感知终端，通过网络实时收集健康感知数据；另一方面，健康服务云平台依据健康知识和健康科学，建立相应的健康评估模型，并基于健康数据的关联分析方法处理感知数据，发现个体或群体的健康问题，及时给出健康促进提醒与指导。

其二，明确线上与线下健康服务的具体内容并合理分工，实现优势互补和良性循环。例如，线上平台依托医养知识库和第三方服务接口，提供健康档案查询、智能健康监护、健康教育、健康咨询、慢性病管理、亲情关爱、上门服务预约、在线健康商品购买等服务；线下则依托社区综合服务中心、社区卫生服务中心、健康管理中心、医疗急救中心

以及其他第三方服务资源，组织实际的长期照护、居家护理、陪同就医、紧急救助和日常生活等服务。

6 结束语

进入现代社会以来，科学技术的发展极大地推动了人类社会的进步，深刻地影响着人们的生产与生活方式。一方面，人类的物质和精神生活日益丰富；另一方面，各种现代文明病快速蔓延，成为人类健康的主要威胁。特别是21世纪以来，全球逐渐进入老龄化社会，导致当前以疾病治疗为中心的医疗服务体系不能完全满足日益攀升的健康服务需求，需要全社会的共同关注和参与。在此背景下，健康感知与计算这一多学科融合的新兴方向越来越多地引起重视，其相关研究的开展需要来自计算机科学、电子信息、健康学、病理学等领域学者的共同努力和协同创新。特别地，对于我国学者而言，还应该借鉴传统文化中人体系统观和均衡论等思想，以推动我国健康感知与计算领域的特色创新。此外，从应用角度而言，需要结合我国人口多、未富先老等具体国情，面向不同人群研发性价比优良的健康感知与服务系统，为构筑可持续的医疗与健康服务体系，提高我国民众健康水平做出贡献。

参考文献

- [1] <http://www.nsf.gov/awardsearch/>.
- [2] http://ec.europa.eu/research/fp7/index_en.cfm?pg=health.
- [3] A Milenkovic, C Otto, E Jovanov. Wireless sensor networks for personal health monitoring: Issues and an implementation. *Comput. Commun.* , 29, pp. 2521-2533, 2006.
- [4] Shnayder V, Chen B, Lorincz K, et al. Sensor networks for medical care. *SenSys 2005*, 5: 314-314.
- [5] Oliver N, Flores- Mangas F. HealthGear: a real-time wearable system for monitoring and analyzing physiological signals. *BSN 2006*.
- [6] Poh M Z, Kim K, Goessling A D, et al. Heartphones: Sensor Earphones and Mobile Application for Non-obtrusive Health Monitoring. *ISWC2009*: 153-154.
- [7] Rai P, Kumar P S, Oh S, et al. Smart healthcare textile sensor system for unhindered-pervasive health monitoring. *SPIE Smart Structures and Materials + Nondestructive Evaluation and Health Monitoring. International Society for Optics and Photonics*, 2012: 83440E-83440E-10.
- [8] Park S, Mackenzie K, Jayaraman S. The wearable motherboard: a framework for personalized mobile information processing[C]. *The 39th ACM annual Design Automation Conference*, 2002: 170-174.
- [9] Pandian P S, Mohanavelu K, Safeer K P, et al. Smart Vest: Wearable multi-parameter remote physiological monitoring system. *Medical engineering & physics*, 2008, 30(4): 466-477.
- [10] M Sung, C Marci, A Pentland. Wearable feedback systems for rehabilitation. *J. NeuroEng. Rehabil.* , vol. 2, p. 17, 2005.

- [11] U Anliker, J A Ward, P Lukowicz, G Tröster, F Dolveck, M Baer, F Keita, E B Schenker, F Catarsi, L Coluccini, A Belardinelli, D Shklarski, M Alon, E Hirt, R Scmid, M Vuskovic. AMON: A wearable multiparameter medical monitoring and alert system. *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 4, pp. 415-427, 2004.
- [12] C Brüser, K Stadlthanner, S Waele, S Leonhardt. Adaptive Beat- to- Beat Heart Rate Estimation in Ballistocardiograms. *IEEE Trans. on Information Technology in Biomedicine*, 15(5), 2011.
- [13] Harada T, Sakata A, Mori T, et al. Sensor pillow system: monitoring respiration and body movement in sleep[C]. IEEE/RSJ International Conference on Intelligent Robots and Systems, 2000: 351-356.
- [14] X Zhu, W X Chen, T Nemoto, Y Kanemitsu, K Kitamura, K Yamakoshi. Accurate determination of respiratory rhythm and pulse rate using an under pillow sensor based on wavelet transformation[C]. The 27th Annual International Conference of the Engineering in Medicine and Biology Society. 2005, 5869-5872.
- [15] M Zanon, M Riz, Gi Sparacino, A Facchinetto, R E Suri, M S Talary, C Cobelli. Assessment of Linear Regression Techniques for Modeling Multisensor Data for Non-Invasive Continuous Glucose Monitoring[C]. The 33rd IEEE Annual International Conference of EMBS, pp. 2538-2541, 2011.
- [16] Wei R, Li X, Im J J, et al. A development of pillow for detection and restraining of snoring[C]. The 3rd International Conference on Biomedical Engineering and Informatics, 2010, 4: 1381-1385.
- [17] Suryawanshi R, Zende A. Electronically Operated Anti- snoring Pillow [C]. The 2nd International Conference on Computer Engineering and Applications, 2010: 626-628.
- [18] Z Chen, J Teo, S Ng, H Yim. Smart Pillow For Heart Rate Monitoring Using A Fiber Optic Sensor. Proc. of SPIE, 2011.
- [19] J Hao, M Jayachandran, P Kng, S Foo, P Aung, Z Cai. FBG- based smart bed system for healthcare applications. *Front. Optoelectron*, 3(1), pp. 78-83, 2010.
- [20] S Šprager, D Zazula. Heartbeat and Respiration Detection from Optical Interferometric Signals by Using a Multimethod Approach. *IEEE Trans on Biomedical Engineering*, 59(10), 2012.
- [21] N Thakor, Y Zhu. Applications of adaptive filtering to ECG analysis: noise cancellation and arrhythmia detection. *IEEE Transactions on Biomedical Engineering*, 38(8), 1991.
- [22] V Somers, M Dyken, M Clary and F Abboud. Sympathetic Neural Mechanisms in Obstructive Sleep Apnea. *J. Clin. Invest.*, 96: 1897-1904, 1995.
- [23] Y Zhu, D Shasha. StatStream: Statistical monitoring of thousands of data streams in real time. VLDB 2002: 358-369.
- [24] V Guralnik, J Srivastava. Event detection from time series Data. KDD 1999: 33-42.
- [25] I Batal, H Valizadegan, G Cooper and M Hauskrecht. A Pattern Mining Approach for Classifying Multivariate Temporal Data [C]. IEEE International Conference on Bioinformatics and Biomedicine, Atlanta, Georgia, November 2011.
- [26] A Wright, T N Ricciardi, M Zwick. Application of Information- Theoretic Data Mining Techniques in a National Ambulatory Practice Outcomes Research Network. *AMIA Annu Symp Proc.*, 829-833, 2005.
- [27] T Shafizadeh, E Moler1, J Kolberg, U Nguyen, T Hansen, T Jorgensen, O. Pedersen, K Borch-Johnsen. Comparison of Accuracy of Diabetes Risk Score and Components of the Metabolic Syndrome in Assessing Risk of Incident Type 2 Diabetes in Inter99 Cohort. *Plos ONE*, 6(7), 2011.
- [28] GA Nichols, EJ Moler. Diabetes incidence for all possible combinations of metabolic syndrome components. *Diabetes Res Clin Pract*, (90): 115-121, 2010.
- [29] RM Conroy, K Pyörälä, AP Fitzgerald, S Sans, A Menotti, G Backer. SCORE project group,

- Estimation of ten-year risk of fatal cardiovascular disease in Europe. *Eur Heart J*, (24), pp. 987-1003, 2003.
- [30] V Tresp, T Briegel, J Moody. Neural-Network Models for the Blood Glucose Metabolism of a Diabetic. *IEEE Transactions on Neural Networks*, 10(5), pp. 1204-1213, 1999.
- [31] Z Wei, K Wang, H Qu, H Zhang, J Bradfield, C Kim, E Frackleton, C Hou, J Glessner, R Chiavacci, C Stanley, D Monos, S Grant, C Polychronakos, H Hakonarson. From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes [J]. *PLoS Genetics*, 5(10), 2009.
- [32] S Palaniappan, R Awang. Intelligent Heart Disease Prediction System Using Data Mining Techniques [J]. *International Journal of Computer Science and Network Security*, 8(8), 2008.
- [33] C Dangare, S Apte. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques [J]. *International Journal of Computer Applications*, 47(10), 2012.
- [34] J Wu, J Roy, W Stewart. Prediction Modeling Using EHR Data: Challenges, Strategies, and a Comparison of Machine Learning Approaches [J]. *Medical Care*, 48(6), 2010.
- [35] D Noble, R Mathur, T Dent, C Meads, T Greenhalgh. Risk models and scores for type 2 diabetes: systematic review. *BMJ* 2011, 343: d7163.
- [36] A Abbasi, L M Peelen, E Corpeleijn, Y T van der Schouw, R P Stolk, A M Spijkerman, A D L van der, K G Moons, G Navis, S J Bakker, J W Beulens. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ* 2012, 345: -5900.
- [37] GC Siontis, I Tzoulaki, K C Siontis, J P Ioannidis. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *BMJ* 2012, 344: -3318.
- [38] S Ahmad, A Tejuja, K Newman, R Zarychanski, A Seely. Clinical review: a review and analysis of heart rate variability and the diagnosis and prognosis of infection. *Crit Care*, 13(6), 2009.
- [39] The Society for Complexity and Acute Illness. <http://www.scai-med.org/>.
- [40] 吕姿之. 健康教育与健康促进 [M]. 北京: 北京医科大学出版社, 2002.
- [41] N J Pender, C Murdaugh, M A Parsons. Health Promotion in Nursing Practice [M], Boston, MA: Pearson, 2011.
- [42] G Dahlgren, M Whitehead. Policies and strategies to promote social equity in health Stockholm. Institute of Futures Studies, 1991.
- [43] A Beattie. Knowledge and control in health promotion: A test case for social policy and social theory. Gabe, J Calnan, M Bury, M(Eds), The sociology of the health service. London, Routledge, 1991.
- [44] L W Green, M W Kreuter. Health promotion Planning: An educational and ecological approach [M]. New York, McGraw Hill, 2005.
- [45] S Consolvo, J Landay, D McDonald. Designing for Behavior Change in Everyday Life. *IEEE Computer*, 42 (6), pp. 86-89, 2009
- [46] I Albaina, van der Mast CAPG, T Visser, MH Vastenburg. Flowie: A Persuasive Virtual Coach to Motivate Elderly Individuals to Walk. *Pervisave Health* 2009.
- [47] A Grimes, V Kantroo, RE Grinter. Let's Play! Mobile Health Games for Adults. *UbiComp'10*, 2010.
- [48] S Munson, D Lauterbach, M Newman, P Resnick. Happier Together: Integrating a Wellness Application into a Social Network Site. *PERSUASIVE'10*, pp. 27-39, 2010.
- [49] R Oliveira, M Cherubini, N Oliver. MoviPill: Improving Medication Compliance for Elders Using a Mobile Persuasive Social Game. *UbiComp'10*, 2010.

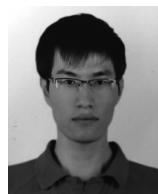
- [50] M Newman, D Lauterbach1, S Munson, P Resnick, M Morris. It's not that I don't have problems, I'm just not putting them on Facebook. Challenges and Opportunities in Using Online Social Networks for Health. CSCW'11, 2011.
- [51] J M Jones, J Nyhof- Young, A Friedman, P Catton. More than just a pamphlet: Development of an innovative computer-based education program for cancer patients. Pat. Edu. Counsel. , 44 (3), pp. 271-281, 2001.
- [52] S Vurgun, M Philpose, M Pavel. A statistical reasoning system for medication prompting. The 9th Int. Conf. Ubicomp, pp. 1-18, 2007.
- [53] T L Hayes, K Cobbina, T Dishongh. A study of medication-taking and unobtrusive, intelligent reminding. Telemed. J. e-Health, 15(8), pp. 770-776, 2009.
- [54] Dongmin Guo, David Zhang, Naimin Li, Lei Zhang, Jianhua Yang. Diabetes Identification and Classification by Means of a Breath Analysis System. Medical Biometrics, Lecture Notes in Computer Science, 6165: 52-63, 2010.
- [55] Lei Liu, Wangmeng Zuo, David Zhang, Naimin Li, Hongzhi Zhang. Combination of Heterogeneous Features for Wrist Pulse Blood Flow Signal Diagnosis via Multiple Kernel Learning. IEEE Transactions on Information Technology in Biomedicine, 16(4): 598-606, 2012.
- [56] David Zhang, Bo Pang, Naimin Li, Kuanquan Wang, Hongzhi Zhang. Computerized Diagnosis from Tongue Appearance using Quantitative Feature Classification [J]. The American Journal of Chinese Medicine(AJCM) , 33(6): 859-866, 2005.
- [57] J Zhang, R Wang, S Lu, J Gong, Z Zhao, H Chen, L Cui. EasiCPRS: Design and Implementation of a Portable Chinese Pulse-wave Retrieval System[C]. The 9th ACM Conference on Embedded Networked Sensor Systems(SenSys) , 2011.
- [58] 孙磊. 健康管理中时序数据挖掘相关问题研究与应用[D]. 硕士学位论文. 清华大学, 2011.
- [59] 吴巧玉. 心血管健康评估表的初步建立[D]. 硕士学位论文. 浙江大学, 2013.
- [60] 李磊. 基于规则引擎的健康评估系统的设计与实现[D]. 硕士学位论文. 上海交通大学, 2013.
- [61] F Guo, Y Lin, S Li, Y Dai. Interval-Valued Cloud Model Based Personal Sub-health Status Diagnosing Prototype System on TCM Syndrome Data. UIC/ATC 2012, pp. 803-810, 2012.
- [62] 杨军, 俞梦孙, 张宏金, 成奇明. 睡眠周期的中医解读[C]. 第十届中国科协年会论文集, 2008.
- [63] 李金林. 健康促进的创新研究[D]. 博士学位论文. 浙江大学, 2011.
- [64] L Tang, X Zhou, Z Yu, Y Liang, D Zhang, H Ni. MHS: A Multimedia System for Improving Medication Adherence in Elderly Care. IEEE Systems Journal 5(4), pp. 506-517, 2011.
- [65] 岳宇. 生物雷达检测技术中心跳与呼吸信号分离技术的研究[D]. 硕士学位论文, 2007.
- [66] 黄莉, 史林, 姜敏. 基于提升算法的低速目标信号提取与生命信号检测应用[J]. 电子科技, 2004, 5: 18-21.
- [67] 史林, 姜敏, 黄莉. 基于谐波模型的生命探测雷达人体状态识别方法[J]. 西安电子科技大学学报(自然科学版), 2005, 32(2): 179-183.
- [68] 王冬雪. 齿痕舌的识别及其与亚健康状态之间相关性的研究[D]. 哈尔滨工业大学硕士论文, 2011.
- [69] 赵云龙. 便携式亚健康监控系统的研究与设计[D]. 西南交通大学硕士论文, 2011.
- [70] 杨军, 俞梦孙, 张宏金, 等. 睡眠呼吸障碍检测技术研究与应用课题组集体专著, 微动敏感床垫睡眠监测系统检测睡眠呼吸事件的原理与判断规则[M]. 人民军医出版社出版, 2011.

作者简介

周兴社 西北工业大学计算机学院教授，博士生导师，主要从事网络化嵌入式计算与普适计算、分布式计算与云计算及其应用研究。陕西省云计算技术工程中心主任，陕西省嵌入式系统重点实验室主任；国务院第六届学科评议委员，中国计算机学会常务理事，中国计算机学会嵌入式系统（微机）专业委员会、普适计算专业委员会副主任委员，中国计算机学会西安分部主席；长期主持国家自然基金重点课题、国家重大专项课题、国家高新技术研究计划课题、国防预先研究课题等。



王柱 工学博士，西北工业大学计算机学院讲师，目前主要从事普适计算与健康计算领域的研究。



倪红波 工学博士，西北工业大学计算机学院副教授，主要研究方向为普适计算、智能系统技术。



王天本 西北工业大学在读博士研究生，主要研究方向为普适计算与健康计算。



林强 工学博士，西北民族大学数学与计算机科学学院副教授，主要研究方向为健康计算与智能辅助技术。



面向智能视频监控的视觉感知与处理

CCF 多媒体专业委员会

傅慧源¹ 黄铁军² 姜育刚³ 李 波⁴ 马华东¹ 薛向阳³ 于俊清⁵ 郑 锦⁴

¹北京邮电大学计算机学院，北京

²北京大学信息科学技术学院数字媒体研究所，北京

³复旦大学计算机科学技术学院，上海

⁴北京航空航天大学计算机学院，北京

⁵华中科技大学计算机科学与技术学院，武汉

摘要

随着视频监控技术的日益成熟和监控设备的普及，视频监控应用日益广泛，监控视频数据量呈现出爆炸性的增长，已经成为大数据时代的重要数据对象。然而由于视频数据本身的非结构化特性，使得监控视频数据的处理和分析相对困难。面对大量摄像头采集的监控视频大数据，如何有效地按照视频的内容和特性去传输、存储、分析和识别这些数据，已经成为一种迫切的需求。面向智能视频监控中大规模视觉感知与智能处理问题，本报告围绕监控视频编码、目标检测与跟踪、监控视频增强、视频运动与异常行为识别等四个主要研究方向，系统阐述 2013 年度的技术发展状况，并对未来的发展趋势进行展望。

关键词：视频监控，目标检测，目标跟踪，视频增强，行为识别

Abstract

With the increasing maturity of video surveillance technologies and popularity of surveillance equipment, video surveillance applications are increasingly widespread. The amounts of surveillance video are showing the explosive growth. In the era of big data, the data of surveillance video has become one of the important data objects. However, due to the unstructured nature of video data, the processing and analysis of multimedia data is relatively difficult. In face of huge video data captured by a large number of surveillance cameras, how to effectively transmit, store, analyze and identify in accordance with the multimedia content and features, has become an urgent need. For the problems of large scale visual perception and intelligent processing in the area of intelligent video surveillance, this report is organized around surveillance video encoding, target detection and tracking, augmented surveillance video together with video motion and identifying abnormal behavior four research directions, and elaborate their development status in 2013 and future development trend outlook.

Keywords: video surveillance, target detection, target tracking, video enhancement, behavior recognition

1 引言

目前，我国大型城市的视频监控摄像头数量通常都在数十万个以上，全国城市已安装的摄像头已经超过 2000 万个。千千万万个摄像头通过互联网连接形成了一张覆盖全球的“视听感知网”，从此人类社会的运行状态都被海量的摄像头采集下来，视频监控已经成为继数字电视、视频会议之后的又一个重大视频应用，而且日益成为“体量”最大的一个视频系统。

根据国际数据公司（International Data Corporation, IDC）的研究报告，2012 年全球各种数据的总量为 2.84ZB，到 2020 年将上升到 40ZB，IDC 称之为“数字宇宙（Digital Universe）”。“数字宇宙”中有分析利用价值的部分是目前热议的“大数据（Big Data）”。IDC 估计 2012 年的数据中“大数据”占 23%，其中一半是监控视频；2020 年这个比例将增长到 33%，监控视频将占 44%，即 2020 年全球监控视频的数据量将达到 5.8ZB，是 2012 年全球数据量的两倍。以北京市为例，若一百万个摄像头都达到高清标准，每个小时产生的视频总量为 $10\,000\text{Gbps} \times 3\,600/\text{h} = 36\,000 (\text{Tb}/\text{h}) = 4.5 (\text{PB}/\text{h})$ 。2012 年存储价格为每 GB 2 美元，北京市存储一小时的监控视频需要 5 600 万元，一个月就是 400 亿。全国 2 000 万个监控摄像头拍摄的数据，存储一个月需要 8 000 亿。

面对大量摄像头采集的海量监控视频和高昂的存储成本，如何对其进行有效地编码、传输、分析和识别已成为当前多媒体领域面临的重大挑战。本文将对 2013 年度的智能视频监控技术做综述性介绍，内容涉及监控视频编码、目标检测与跟踪、监控视频增强、视频运动与异常行为识别。

2 国际研究现状

2.1 监控视频编码

视频压缩又称视频编码，其目标是通过各种技术手段大幅度降低视频码率。高清晰度视频在不压缩的情况下码率约 1.5Gbps，即使是今天的带宽条件，传输这样的一路视频依然成本很高。从 1952 年贝尔实验室 Cutler 等人进行脉冲编码调制（Differential Pulse Code Modulation, DPCM）技术研究算起，视频编码技术的研究历史已经 60 多年。20 世纪 80 年代，为了数字电视和视频通信的需要，国际标准组织开始综合已有技术成果制定视频编码标准，形成了以块为单元的预测加变换的混合编码框架，并相继出台了国际电信联盟（ITU-T）H.261/2/3/4 视频编码建议和国际标准化组织和国际电工技术委员会

(ISO/IEC) 的 MPEG (Moving Picture Experts Group, 运动图像专家组) – 1/2/4 视频编码标准, 其中 1994 年出台的 MPEG-2 标准在数字电视领域得到了广泛采用, 压缩比可以达到 75 倍, 即可以把高清视频压缩到 20Mbps 左右。国际电信联盟 1995 年出台的 H. 263 标准也是同一时代的技术, 在视频会议领域得到广泛应用。第一代数字视频监控系统就借用了 MPEG-2 或 H. 263, 有的系统还进行了一定的简化。

2003 年第二代视频编码技术标准出台。国际的标准为 ITU-T H. 264 和 ISO/IEC MPEG-4 AVC, 二者为同一套技术, 两个渠道出版^[1,2]。面对国际标准背后高昂的专利费问题, 我国在国际标准出台约一年之后制定了具有自主知识产权的国家标准, 并经过芯片实现等产业化验证后, 于 2006 年 2 月颁布为《信息技术先进音视频编码第二部分视频》国家标准 (国标号 GB/T 20090. 2-2006, 通常简称为 AVS 视频编码标准)^[3]。4 个月后, 微软主导的 VC-1 视频编码标准由美国电影电视工程师协会 (The Society of Motion Picture and Television Engineers, SMPTE) 颁布为行业标准。这三个标准通常被称为第二代视频编码标准, 其编码效率均比第一代翻了一番, 压缩比达到 150 倍左右, 即可以把高清视频 (在质量达到广播要求的情况下) 压缩到 10Mbps 以下。第二代标准在数字电视和视频通信领域得到应用后, 也很快被视频监控系统所采用, 目前基于 IP 的网络视频监控系统, 很多都采用 H. 264 标准, 但厂商为了降低成本, 往往会把标准中较为复杂的编码工具剪裁掉, 而不同厂商剪裁的方式又各不相同, 因此虽然都号称基于 H. 264 标准, 但是产品还是不能互通。

2013 年上半年, 第三代视频编码国际标准 ITU-T H. 265 和 ISO/IEC HEVC (High Efficiency Video Coding, 高效视频编码) 颁布, 编码效率比 H. 264 又提高一倍。下面将从预测、变换、量化、扫描和熵编码等方面介绍 HEVC 标准采用的新技术^[4]。

编码、预测和变换单元: 在 H. 264/AVC 中, 一个 16×16 的宏块可划分成从 16×16 到 4×4 共 7 种不同帧间预测尺寸模式和 16×16 、 8×8 、 4×4 三种帧内预测尺寸模式。而在 HEVC 中, 编码的基本单元的大小从 H. 264 与 AVS 的 16×16 宏块扩展到了 64×64 编码树单元 (Coding Tree Unit, CTU) 的超大宏块以便于高分辨率视频的压缩。为了更有效地进行数据的压缩, HEVC 对编码单元采用了更加灵活的方式对 CTU 进行表示: 编码单元 (Coding Unit, CU)、预测单元 (Predicting Unit, PU), 变换单元 (Transforming Unit, TU)。

帧内预测: 单一图像内通常都有很高的空间相关性, H. 264/AVC 和 AVS 都引入了空域的帧内预测技术。相比于 H. 264 中的 8 种方向性预测模式, HEVC 提供的方向预测具有更高的精度。HEVC 中的帧内预测提供多达 35 个预测模式, 其中包括 33 个方向性预测 (如图 1 所示) 和两个非方向性预测模式 Planner 和 DC。

多帧参考: 在 H. 264/AVC 中, 可采用多个参数帧的运动估计, 即在参考图像缓存中存有多个重构编码帧, 当前单个编码宏块可以从这些帧中选择一个更好的参考宏块, 从而获得比单帧参考更好的编码效果。HEVC 标准中, 除了支持 H. 264 中已经拥有的层次 P 帧、B 帧编码结构之外, 通过可配置的参考图像集的方式支持了极为灵活的参考帧选择方式, 通过使用参考帧列表的合并提高非低延时情况下普通 B 帧的编码效率。

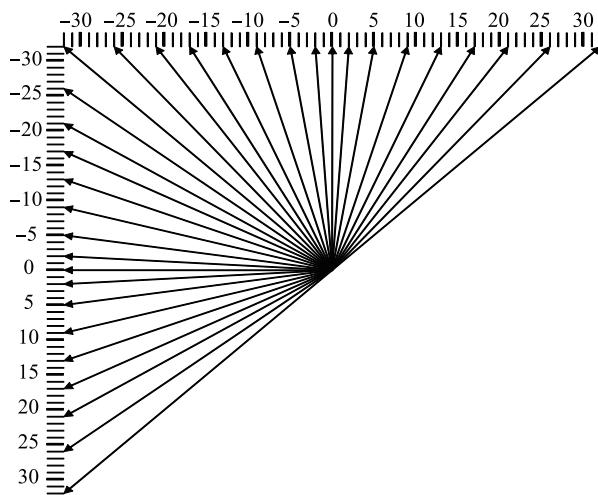


图 1 HEVC 定义的 33 个帧内预测方向

运动矢量预测：运动矢量预测（Motion Vector Prediction, MVP）利用相邻块之间运动矢量的相关性来减少编码运动矢量所占用的码率。与 H.264/AVC 和 AVS 使用相邻块的运动矢量直接导出算法不同，HEVC 采用竞争机制的 MVP，即所谓的高级运动矢量预测（Advanced Motion Vector Prediction, AMVP）。在 AMVP 技术中，待选的预测运动矢量（Predicted Motion Vector, PMV）来自左侧，上侧或者时域对应位置的 PU，编码器从这三个运动矢量待选列表中选择其中两个（如果可用的 PMV 不够两个，那么用 0 运动矢量补足），计算它们的率失真代价，选择率失真代价最低的 PMV 作为最终的 PMV，并将其索引号进行编码。

高像素精度运动补偿：运动矢量的精度是提高预测准确度的重要手段之一。H.264/AVC 中采用了 1/4 像素精度的运动补偿，其中半像素位置采用 6 拍滤波，1/4 像素位置采用双线性插值，对于色度则是 1/8 像素双线性插值。HEVC 对于帧内预测采用 1/32 像素精度的插值，对帧间预测采用 8 拍亮度 1/4 像素插值，对色度采用 4 拍 1/8 像素插值。对于 HEVC 的编码器内部，增加了像素比特深度以及高精度的双向运动补偿技术。此外，在插值过程中，HEVC 采用一种基于离散余弦变换的插值滤波器（Discrete Cosine Transform-Interpolation Filter, DCT-IF）生成分数像素。DCT-IF 首先利用 N 个整数像素的像素值，计算它们的 DCT 系数，然后利用这些系数生成一条由 DCT 基构成的光滑曲线，分数像素的取值则可以根据这条光滑曲线的表达式计算得到。

多种帧间预测模式：H.264/AVC 支持丰富的帧间预测模式，包括前向、后向、双向和直接/跳过模式，每个宏块划分的子块都可以从上述模式中选择各自最优的方式。HEVC 增加了广义 B 帧（Generalized P picture replaced by B picture, GPB）预测方式取代低时延应用场景中的 P 预测方式，如图 2 所示，前向和后向参考列表中的参考图像都必须为当前图像之前的图像。此外，HEVC 使用了合并（Merge）模式来取代和扩展了 H.264/AVC 中的直接跳过模式。

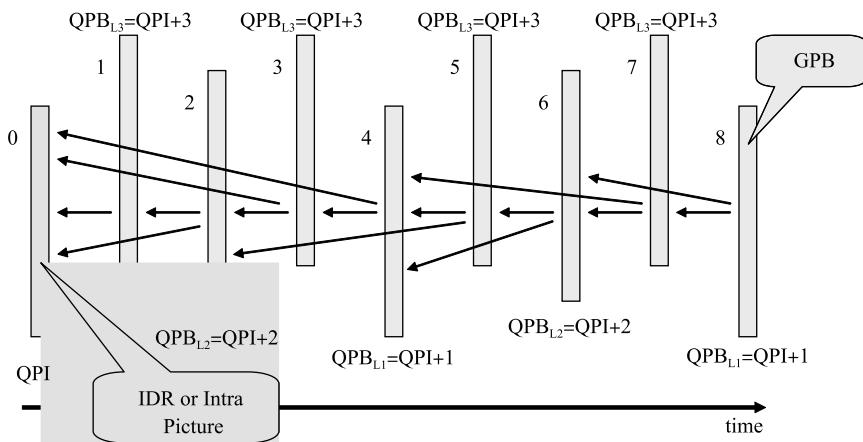


图 2 HEVC 中无延迟编码时 GPB 帧的使用

低复杂度整数变换及无除法量化：H. 264/AVC 采用近似于 DCT 的整数变换，大大降低变换的复杂性，可以只通过移位和加法来实现整个变换过程，不存在正反变换误差问题。HEVC 中，采用了一种自适应的变换技术机制，对于帧内预测和帧间预测使用不同的残差四叉树变换（Residual Quad-tree Transform，RQT），在帧内编码中变换单元的大小应不大于预测单元的大小，而在帧间编码中，变换单元的大小不一定小于预测单元的大小，但一定小于等于编码单元的大小。

环路滤波：基于块的视频编码会产生编码块效应，去块效应滤波器是提高视频质量的有效方法之一。HEVC 的完整环路滤波过程包括去块滤波和样点自适应补偿（Sample Adaptive Offset，SAO）两个环节。去块滤波在 H. 264 的去块滤波技术基础上发展而来，但为了降低复杂度，取消了对 4×4 块的去块滤波。使用 SAO 技术，按照递归的方式将重构图像分裂成 4 个子区域，为减少预测残差，根据其图像像素特征给每个子区域选择一种像素补偿方式。

基于上下文的适应性变长/算术编码：与以往传统标准中多采用固定码表变长编码的方式相比，H. 264/AVC 引入全新的上下文自适应变长编码（Context Adaptive Variable Length Coding，CAVLC）以及上下文自适应算术编码（Context Adaptive Binary Arithmetic Coding，CABAC），这两种编码方法充分挖掘编码元素的上下文相关性，根据上下文选择合适的不同模型进行编码，进一步提高了编码效率。为了解决 CABAC 的吞吐能力问题，HEVC 使用基于语法元素的并行 CABAC 编码方案（Syntax-based context-adaptive Binary Arithmetic Coding，SBAC）。

HEVC 被监控行业寄予厚望。但是，视频编码标准的更新换代和压缩效率的提高，都是以更高的计算复杂性换来的，压缩效率提高一倍，计算复杂度往往要提高五倍甚至更多，从而导致编码器价格居高不下。根据文献[4]的分析，HEVC 解码器复杂度与 H. 264 相差不大，但是编码器复杂度是 H. 264 的数倍以上，因此 HEVC 实时编码器开发还需要一段时间。对于电视广播来说，每个频道一台编码器就可以服务亿万用户，因此编码器复杂度高、价格高不是大问题。但是，视频监控与数字电视恰恰相反，解码器需

求不多（很多视频可能从未解码查看过），但每个摄像头都需要一个编码器，这就要求在提高压缩效率的同时，编码算法复杂度应该保持较低的水平。

与监控视频编码相关的研究方面，Musmann 提出打破基于分块的编码而以对象为单位编码形状、颜色信息和预测残差^[5]。在此基础上，文献[6~8]进一步为监控视频设计了这种面向对象的方法。随着基于对象的标准 MPEG-4 制定^[9]，Francois 等继承了基于背景建模和背景剪除的更准确对象检测、跟踪、识别和分割成果，提出了一种基于准确区域分割的面向对象的视频编码方法^[10]。针对监控视频，为了实现更高效的存储，文献[11]关注于对象分割，忽略监控视频背景。为了使背景质量不至于太差，文献[12, 13]提出在混合编码框架下编码背景残差，并且和前景表示残差一起进行编码。

2.2 目标检测与跟踪

运动目标检测与跟踪作为视频运动对象分析的核心内容，涉及计算机视觉、机器学习、模式识别、人工智能、概率统计和随机过程等领域，并广泛应用于空防预警、生物组织运动分析、气象云图预报、交通监控管理、安防视频监控、高清电视频带压缩、基于内容的检索等方面。一个典型的运动目标检测与跟踪应用场合是视频监控，在该应用下，具有这两项功能的系统通过分析视频序列，快速、准确地检测到单个或者多个运动目标，计算出目标在每帧图像上的位置，实现对目标速度的估计，同时自动发送控制指令使摄像机自动跟踪目标。

20世纪70年代以来，视频运动目标检测和跟踪一直受到了广泛关注。近年来，随着视频摄像头的大量应用，造成了海量视频数据越积越多，随之带来的“信息多、用不好”问题推动了视频运动目标检测跟踪理论与应用的不断发展。其中所产生的背景减（Background Subtraction, BS）、高斯混合模型（Gaussian Mixture Model, GMM）、光流（Optical Flow, OF）、均值漂移（Mean-Shift, MS）、粒子滤波（Partical Filter, PF）、主动轮廓模型（Active Contour Models, ACM）等方法，也极大地推动了计算机视觉相关技术的发展。

监控视频具有受场景噪声大、目标类型多样、场景和目标状态多变等特点影响，而现有的运动对象分析方法在有背景噪声、无意义运动干扰时难以准确检测到慢速或者小的运动目标；在目标发生尺度缩放、遮挡、跳变、消失，以及光照变化、干扰物体时难以正确跟踪。因此，需要突破像素级处理和中低层特征分析方法的局限，结合时域、空域、频域，以及不同尺度特征的互补性，提高运动目标检测的准确性，在构建特征提取新算子的基础上根据目标状态选择有效目标进行跟踪，提高复杂场景下视频运动目标跟踪的准确性和适应性。

虽然视频运动目标检测跟踪及其应用研究已取得了一定成果，但是作为一项多学科交叉融合的前沿课题，且实际应用中需要适应以下复杂场景：1) 树枝叶晃动等背景扰动，雾气、沙尘等噪声干扰，光照变化和阴影，摄像机无规律运动；2) 目标小且运动速度慢，目标存在被遮挡、形变、姿态变化、外观变化；3) 相似物干扰，使得要实现一个

准确性高、适应性强、实时性好的视频运动目标检测与跟踪系统成为计算机视觉中一个公认的难题。从最近几年计算机视觉领域顶级国际会议（CVPR、ECCV）和国际期刊（PAMI、TIP）上发表的论文^[14~22]来看，视频运动目标检测和跟踪仍然是目前计算机视觉领域中研究的热点问题。由于运动目标检测与跟踪是两个相对独立的过程，下面就其中涉及的技术发展现状分别进行阐述。

2.2.1 运动目标检测

运动目标检测可分为变化检测、运动检测和特征检测三类。变化检测得到变化点或变化区域；运动检测得到目标的运动矢量，它们主要利用了视频的时域信息；特征检测利用视频的空域信息，即利用图像中目标特征进行检测。

(1) 变化检测

变化检测通过比较帧间差别获取变化区域，该变化区域被认为是由目标运动造成的。变化检测包括相邻帧差法和背景减法。相邻帧差法将视频序列中相邻两帧差的绝对值大于阈值的像素点标识为运动点，计算简单、检测速度快，且不受光照缓慢变化的影响，适合摄像机无运动、噪声较小的情况，但检测结果受目标运动速度影响大，容易造成空洞。背景减法在训练阶段建立背景模型，进而和当前帧比较提取运动目标，常用的有时域差分法、均值（中值）阈值法、高斯混合模型法、线性预测法、特征背景法、归一化块相关法等。根据背景模型建立的空间尺度，前四种方法在像素级上处理，归一化块相关法在区域级上处理，特征背景法在帧一级处理。WallFlower^[23]提出结合像素级、区域级和帧级的背景维护，在像素级用维纳滤波器预测期望的背景图，在区域级填充前景，在帧级对背景图进行全局更新。此外，近年来有学者提出字典学习^[24,25]的方法检测运动目标，这种方法不是基于像素点更新，而是直接全局更新，可以在一定程度上解决光线敏感的问题。背景减法虽然得到了广泛应用，但是还存在一些固有的问题，如目标太小时容易被背景淹没，检测效果受场景中遮挡、光线变化、背景扰动等影响，容易出现误检。此外，背景运动下的检测结果依赖于预处理效果，即使采取措施来补偿背景运动，通常也会产生大量虚假目标。但是作为一种效率较高的算法，背景减法仍然广受关注和应用。

(2) 运动检测

运动检测利用目标与背景表观运动模式不同检测运动目标，具有代表性的方法为光流法，当摄像机与场景目标间有相对运动时所观察到的亮度模式运动称为光流。光流法又分为微分法、匹配法、能量法、相位法和小波法。微分法假定图像在空间和时间上是连续的，利用图像强度的时空导数来计算每一像素点的速度矢量；匹配法在相邻图像中查找像素（块）之间的对应关系，两幅图像对应像素（块）之间的位移即是所求的光流；能量法基于速度调谐滤波器的输出能量计算光流，将光流估计转化为时空能量与频率空间的最小二乘拟合问题；相位法根据与带通速度调谐滤波器输出中的等相位轮廓相垂直的瞬时运动来定义分速度；小波法利用小波变换多尺度多分辨率框架，在不同的频带上构造特征匹配计算光流。光流法适用于摄像机运动下的运动目标检测，但是目前光流计算的普适性、准确性和实时性为亟待解决的问题，更多光流方法的评估可参考文献[26]。

(3) 特征检测

边缘、纹理等对光照变化不敏感，这些空间信息可以提高运动目标检测的准确性，特征检测正是通过建立图像帧空间特征之间的对应关系来检测运动目标。在某些特定应用中，例如运动目标为人脸或红外视频中，可以利用肤色、形状、区域特性等目标特有的特征进行检测^[27~29]。更一般地，可以提取图像中的显著区域（Salient Regions），例如，Cheng^[30]等提出利用全局对比度差异和空间一致性提取显著区域，Li 等^[31]提出自底向上的显著性检测机制，采用图像频谱和低通高斯核作卷积检测显著区域；作为一种局部特征提取算子，尺度不变特征转换（Scale Invariant Feature Transform, SIFT）^[32]特征具有较强的抗噪能力和匹配能力，Liu 等^[33]基于 SIFT 流实现了变化场景图像之间的配准；Rublee 等^[34]提出快速二元特征描述子，该特征描述子具有旋转不变性，且对噪声具有很好的适应性。

检测到特征并完成特征匹配后，需要在图像帧间建立特征之间的对应关系，常用的方法有公式法和最优化法。公式法适用于刚体目标，其运动基本上为平移、尺度缩放、旋转和扭曲，因此可以用仿射变换公式表示，利用二维图像特征对应位置求解运动参数方程。而最优化方法适用于非刚体目标，这些目标存在一定变形，可利用多次迭代求取最优解。

特征检测能够在一定程度上抵抗噪声和光照突变等影响，但是，在提取特征时一般要求根据先验知识确定应提取什么样的特征，而建立特征对应关系时公式法要求的目标为刚体在实际情况中往往不能满足，最优化方法可能不收敛或者陷入局部最优解。因此，实际系统中往往结合时域和空域信息提高检测的准确性。

2.2.2 运动目标跟踪

运动目标检测得到运动点或运动区域，通过连通区域判断和区域聚类可以得到若干个运动目标，进而开始跟踪。运动目标跟踪利用目标具有的可区分性特征，在连续帧中完成运动目标的匹配，实现视频每一帧中的目标定位。

(1) 目标表观模型

在跟踪过程中，目标通常用表观模型表示，常见的有点模型、几何形状模型、边缘轮廓模型、组合形状模型和骨架模型。其中，点模型采用一个点或者多个点表示目标，适合目标较小的情况；几何形状模型采用矩形或者椭圆等简单几何形状表示目标，适合目标本身形状规则的情况，如人脸一般用椭圆表示；边缘轮廓模型中定义边缘为目标的边界，轮廓是边缘内的区域，适合表示复杂的非刚体目标；组合形状模型将目标分为若干部分，每部分用一个几何形状模型表示，各部分通过“关节”连接，例如，人体 = 头 + 躯干 + 手 + 腿 + 脚，各部分之间的关系依靠运动模型控制，比如关节角度等；骨架模型是对目标提取骨架，常用于表示行人等具有关节结构的目标或刚体目标。表观模型限制了运动的类型，例如，如果目标被表示为一个点，则使用平移变换表示目标的平行移动；如果目标被一个几何形状（如椭圆）表示，则适合使用参数运动模型，目标运动通常采用平移、仿射、投影变换；对非刚体目标，边缘轮廓模型^[35]更为适合，可采用更复杂的运动模型描述形变等。

另一方面，按照建模方式目标表观模型又分为生成模型（Generative Model）和判别模型（Discriminative Model）。生成模型试图构建一个紧凑的外观模型来描述可能的目标外观变化，而判别模型旨在将目标外观与背景外观区分开。最常用的生成模型是基于模板和子空间的目标表观建模，如 Mei 等^[36]将稀疏表达引入视觉跟踪的表观建模中，提出了一种基于稀疏表达的在线目标表观建模方法，实现在噪声环境和遮挡情况下的自适应目标跟踪。判别模型将跟踪视为分类问题，其基础是 Avidan^[37]提出的集成学习框架，该框架采用 AdaBoost 算法在线训练获得前景/背景两类分类器。近年来，研究人员将多实例学习的思想引入到目标跟踪分类器训练过程中^[38]，取得了较好的结果。

由于目标自身运动，需要在跟踪过程中对目标表观模型进行更新以便跟踪算法能根据当前场景使用最新的模型。如何在有效避免跟踪漂移的同时还保持对外观变化的自适应能力，是表观模型研究的一个重点。

（2）目标搜索策略

视频运动目标跟踪中常用的目标搜索策略有全局搜索、最优化搜索和随机搜索三类。由于要在目标状态空间中逐点搜索，导致全局搜索方法的跟踪效率较低，尤其是目标特征维数较高时更严重。为了提高视觉跟踪的速度，一些研究者提出了最优化目标搜索方法，基于梯度的最优化目标搜索、特别是 Mean-shift 算法是其中的典型代表，该类方法以相似性度量函数作为目标函数，在每一帧图像中利用最优化算法迭代搜索目标。但是，该方法要求相似性函数对目标的状态参数可微，且易于收敛到局部极值，对目标快速运动和被遮挡鲁棒性较差。一些研究者提出了随机目标搜索方法，粒子滤波是其中的典型代表，由于它能够处理非线性和非高斯问题，并且具有简单、灵活和易于实现等特征，得到了广泛应用，但是也存在粒子退化和运算速度慢等问题。

目前，大多数目标跟踪算法都基于运动连续性假设，使得目标搜索可以在较小的状态空间中完成。然而，在许多复杂场景中（低帧率、多摄像机接力跟踪、目标突变运动），运动连续性假设并不成立，多尺度和多层次采样是处理这类问题常用的方法。文献[39]通过对未知高维状态空间进行分级并用粒子滤波在其中逐级搜索，以较低复杂度得到了相对准确的跟踪结果。这种由粗及细的分层思想提高了处理速度，但是一旦某层目标丢失，后一层就很难再恢复跟踪，因而对滤波方法和特征选择都提出很高要求，而且当层数增大后整体效果反而可能下降。

2.3 监控视频增强

在实际的监控环境中，由于摄像头的能力有限以及监控环境的复杂多变，视频本身常常含有噪声，质量不佳。最为典型的是受光照条件不足或者过度，雾、霾、雨雪、沙尘等恶劣天气条件等不利情况的影响，造成图像的质量下降。为此，在监控系统中通常需要对获取的监控视频进行各种形式的处理用以改善视频的质量。下面将从通过去雾、去夜、去雨雪、去模糊和超分辨率等五个方面对视频质量增强的国际研究进展进行阐述。

2.3.1 去雾

图像去雾技术是通过一定的手段去除图像中雾的干扰，从而提高图像的质量，以便于得到满意的视觉效果并获取更多有效的图像信息。对图像去雾技术的深入研究，恢复图像颜色、对比度，复原景物细节信息等处理，对减少交通运输、室外监控、侦查、导航、遥感遥测等户外成像系统对天气条件的限制，提高其工作的可靠性和稳定性具有重大的应用价值。

(1) 基于图像处理的方法

图像增强是图像处理领域中的一个传统的话题，也是一个比较活跃的研究领域。从处理方法上分为空间域和频率域两种方法；从处理区域上分为全局增强和局部增强；依据处理对象的不同，又可分为灰度图像处理和彩色图像处理。国际上基于图像处理的图像去雾方法包括基于滤波的方法^[40]和基于色彩特性的方法^[41]等。文献[41]提出了一种基于 Kalman 滤波的雾天车载图像恢复算法，介绍了一种自动判决的雾天退化模型，通过使用道路标记估计出雾天降质图像消失点，并实现了雾天退化模型所有参数的自动判决，通过对模型应用 Kalman 滤波获得了去雾后的图像。文献[42]提出综合应用动态范围压缩、色彩恒常和色彩表现理论，通过逼近良好天气的色彩逼真度来恢复恶劣天气降质图像。

(2) 基于物理模型的方法

国际上关于基于物理模型的图像去雾算法发展较早，文献多数是以大气散射模型为基础，通过简化模型或应用场景几何和深度信息来恢复降质图像对比度，包括基于场景深度的方法^[42]、基于大气散射模型的方法、基于先验信息^[43]等。文献[43]利用已知的 3D 模型获取景深，从而复原雾天图像，与大气散射模型相结合，最终实现雾天图像的复原。文献[43]通过统计发现无雾图像相对于有雾图像具有较高的对比度，进而利用最大化复原图像的局部对比度来达到去雾的目的。

2.3.2 去夜

在低照度环境中（如夜晚环境），光线问题造成的图像对比度下降、噪声、过曝和晕环效应等都会使得监控视频质量严重退化，低质视频中往往缺失了目标的形状、亮度和高维纹理信息，这些都给夜晚环境下的监控分析带来了很多困难。夜色去除工作能够将低照度图像近似无损地转换为白天图像，可以为很多智能监控中的任务带来便利，如特征提取、检测分类、车牌识别、目标跟踪和事件分析等。目前，随着智能视频监控系统的广泛应用，这部分工作已经得到了研究者的关注。一般而言，夜晚图像增强需要解决以下三个问题：

- 1) 图像信息的完整性。在颜色和亮度方面，经过增强的图像需要保持原有颜色的同时又要提高亮度。
- 2) 关键区域的质量增强。低质视频增强的目的之一就是图像边缘增强，在诸如前景目标、人脸等感兴趣关键区域需要进行增强或者复原。
- 3) 夜色建模。人们在照度良好的环境中感知到的图像信息往往由于夜晚环境噪声而变得复杂，因此，去夜并没有公认的数学模型。

目前，在世界范围的很多高校和科研机构都开展了夜色环境图像增强方面的研究。例如，麻省理工学院媒体实验室和美国卡内基梅隆大学机器人研究院等。另外，工业界也对低照度图像增强领域给予了很多关注，主要包括 Adobe 公司和微软研究院等。

在图像增强算法中，直方图均衡是一种最常见的方法^[44]。直方图均衡化通常用来增加图像的全局对比度，但是传统的直方图算法的缺陷是不能保留原始图像的亮度信息且处理中会发生图像细节失真现象。为了对夜间视频进行监控，Raskar 等^[45] 和 Jing 等^[46] 都采用图像融合的方法来解决夜间图像的低亮度、低对比度问题。其中，Raskar 提出了一种基于梯度域将白天背景图像和夜间图像融合的方法，Jing 提出了一种应用于夜间视频监控系统的图像增强和图像融合的算法。文献[47]中，利用基于局部小波系数的对比度增强方法对低照度图像进行增强，该算法能够进行对比度的自适应调整，因此灵活度较强。

2.3.3 去雨雪

视频中雨雪天气去除问题的研究始于 20 世纪 90 年代末，目前该问题的基本解决流程已基本确立，主要分为雨雪颗粒检测和雨雪颗粒去除两个步骤。研究人员已提出了多种对于雨雪去除的量化指标，用于评价雨雪检测和去除的效果。

目前，国际上关注于计算机视觉与天气研究的团队主要包括美国哥伦比亚大学计算机视觉实验室和美国卡内基梅隆大学机器人研究院的光照与图像实验室。哥伦比亚大学研究团队主要从大气物理角度对各种天气进行分析和建模，对于雨的物理模型和运动模型进行了深入的研究，基于这些模型对视频中的雨颗粒进行检测和去除。哥伦比亚大学视觉实验室（Computer Vision Laboratory of Columbia University, CAVE）团队，对雨的物理性质和运动特征进行了分析和建模，并提出了基于形状模型和运动模型的雨滴识别方法，取得了很好的效果^[48]。文献[49]中首次提出了在频率域对雨雪颗粒进行建模和分析，并进行准确检测，可对视频中的雨雪进行去除和增强。文献[50]提出了一种基于混合高斯模型的方法，该方法先利用雨线的大小和亮度获取候选滴，通过混合高斯模型来估计方向。该方法可以处理动态背景和摄像头移动的情况，但是只能适合于可以明显区分雨的情况。

2.3.4 去模糊和超分辨率

去模糊和超分辨率具有很高的应用价值，已经成为国际上图像重建领域极为活跃的研究课题，而且，超分辨率重建方法花费少、成本低引起了国内外的广泛研究。目前，超分辨率算法可以分为：基于重建的方法、基于学习的方法及重建与学习相融合的方法三大类。

(1) 基于重建的方法

基于“重建约束”的超分辨率方法，通常在贝叶斯框架中，嵌入某种约束的光滑先验模型，使“病态”的超分辨率方程正则。基于重建的方法根据处理领域的不同又分为频域和空域的超分辨率算法。基于频域的重建超分辨率（Super Resolution, SR）方法，利用傅里叶变换之间的离散或连续平移、混叠性质估算傅里叶变换系数来估计重新得到的重建的高分辨率（High Resolution, HR）图像。该方法简单可行，缺点是无法有效利

用先验知识，能处理的运动模型也很有限。此后，人们对这一领域提出空域实现方法，在图像像素的尺度上，对图像像素点的变换、约束进而改善图像的质量，主要分为以下三种方法：

1) 概率论方法 (Probability Methods)。根据最大后验概率估计，随机信号可以是加性噪声、低分辨率 (Low Resolution, LR) 的观测图像和理想高分辨率图像。最大后验概率就是在观测到 LR 视频序列的前提下，使 HR 图像的后验概率达到最大。目前，常用的先验概率模型有高斯模型 (Gauss)、泊松 (Poisson) 模型、马尔可夫 (Markov) 模型和吉布斯 (Gibbs) 模型。UweSchmid^[51] 提出集成的贝叶斯框架来统一去非盲去模糊 (non-blind deblur) 和噪声估计，利用贝叶斯最小均值误差进行去模糊。Cho^[52] 提出在最大后验概率 (Maximum a Posteriori Probability, MAP) 估计框架下利用 Radon 变化来估计图像的模糊核达到图像去模糊化的目的。

2) 集合论方法。解决超分辨率图像复原问题，凸集投影方法是目前的主流方法之一。根据集合理论方法，要重建图像可以作为用一些约束凸集进行公式化表征的交集内的一个点。而凸约束的特征主要包括能量有界、平滑和正定等。

3) 小波变换。小波变换思想是将信号在时域或者频域同时具有良好的局部化性质，可以聚焦图像的任何细节，国内外许多学者将它应用于超分辨图像重建领域中。相对于空域或者频域方法，具有较小运算量并且迭代速度快等优点。

(2) 基于学习的超分辨率复原方法

基于学习的超分辨率复原方法的核心思想是学习样本集合中的低分辨率和高分辨率图像，得到两者之间的对应关系。处理低分辨率图像率列就是将高分辨率和低分辨率图像对应关系作为先验条件进行优化。

(3) 基于重建和基于学习相融合的超分辨率复原方法

近些年来，有学者充分利用了基于重建和基于学习两者之间的优势，提出基于两者相融合的方法。Kim 等^[53] 首先将低分辨率的图像采用插值的方法将其变为需要的大小，然后利用核边界回归进行逐片回归提取出一组候选图像集合。Glasner^[54] 利用单帧图像本身存在的相似性，如尺寸大小相同的图像块之间的相似性，使用传统的基于重建的超分辨率算法处理相似的尺寸相同的图像块，而多个相似的尺寸大小不同的图像块可以使用基于样本的超分辨率方法，从而提出了单帧图像的基于重建和基于学习相融合的超分辨率重建方法。Shahar^[55] 提出从单个视频时空超分辨率进行基于重建和基于学习融合的超分辨率重建方法，效果很好。

2.4 视频动作与异常行为识别

在大量应用需求的推动下，基于智能分析的视频监控技术引起了广泛关注，其中包括监控视频中对象的动作识别和异常行为识别等。例如：在公共安全监控领域，可用动作识别方法检测场景中的危险动作，如行人丢置不明包裹等；在家庭监控领域，可用来监测家中独居老人的行为，例如是否发生了意外摔倒等。

国外许多高校与研究机构在视频监控方面的研究起步较早，积累了大量理论基础与实践经验。对不同种类的动作识别技术有很多种不同的划分方法^[56~59]，本文主要针对监控视频中的动作对象是属于“已知的特定动作”还是属于“未知的异常行为”这两种不同的情况，分别阐述它们的发展现状。

2.4.1 特定动作识别

特定动作识别指在视频中对于事先指定的若干种动作进行识别的问题。例如，对于监控画面中人物的行走、奔跑、站立等常见动作进行识别都属于该范畴。由于动作的类别明确，一般将这一类识别过程看成是一种对数据进行分类的问题，即：对于使用不同方法提取到的动作描述特征，采用合适的机器学习算法对其进行分类。根据使用的分类模型本身是否有考虑到时变信息，可以进一步将相关技术分为非时序模型与时序模型两大类。

(1) 非时序模型

非时序模型的典型代表包括常见的支持向量机（Support Vector Machines, SVM）、最近邻（Nearest Neighbors, NN）和各种提升（Boosting）策略等。使用这些算法可以直接对提取到的动作特征进行分类。

由于这些分类方法均是比较成熟的机器学习技术，所以在确定了合适的分类模型后，许多工作都将创新的重点放在了如何选取与设计更有效的动作描述特征来帮助提升最终的识别效果。其中，许多特征的提取都需要预先确定运动的区域，这就要用到的物体检测与跟踪方法。这些特征有些是全局特征，有些是局部特征；有些特征是基于连续的时间段抽取，有些则是基于动作发生的某些时刻的关键帧提取的。在送入分类器前，常将一些不固定长度的描述特征转换成统一长度的词袋特征（Bag of Words, BoW）。此外，在特征获取方面，得益于近年来诸如 Kinect 等硬件设备的发展，一些原始视频在采集环节就额外增加了描述深度的信息^[60,61]，这些信息对于动作识别有很高的价值。

目前，基于 SVM 等传统的分类算法仍被广泛地用来对动作描述特征进行分类。很多工作融入了场景中的上下文信息。例如，在 Lan 等^[62]将某个人与他周围的人的特征进行合并送入 SVM 中对该人的动作进行分类；文献[63]和[64]均在特征描述中加入了人与其他物体交互的上下文信息，之后同样用 SVM 进行分类。

总体来说，这类非时序的分类方法有很好的通用性，对于从监督样本上提取到的不同含义的特征都能灵活适用。但是，这些算法属于静态的模型，它们本身不具备模拟先后产生的动作特征之间的时序关联的能力。

(2) 时序模型

顾名思义，时序模型的主要特点是在建模的过程中考虑了动作发生的先后顺序。对于时序模型，除了传统的动作模版匹配方法外，大量近期工作均采用状态空间模型。

隐马尔可夫模型（Hidden Markov Model, HMM）作为一种典型的状态空间模型，被广泛应用在自然语言理解与语音识别等领域，它也是目前为止在视频监控中被使用最多的时序模型之一。HMM 算法利用模型中隐含状态的转变来模拟人体动作的不同阶段的变化，于是动作识别的过程被转化成了找出以最大概率产生观测序列的 HMM 模型所对应

的动作类别的过程。有大量工作在原始 HMM 的基础上针对不同的问题进行了各种改进。为了模拟视频中人体不同运动部分的独立并发运动，提出了更为泛化的动态贝叶斯网络 (Dynamic Bayesian Network, DBN)，这类方法可以进一步模拟人体不同部位的各自动作、画面中不同人的各自运动、人与物之间的各自运动，还有不同视角下的观测到的动作等。Wang 等^[65]在 DBN 的模型中加入了物体与场景的时空上下文的信息来帮助最终分类。

HMM 是一种生成模型 (Generative Model)，在它的基础上还发展出了另一种判别式的条件随机场 (Conditional Random Field, CRF) 模型。在 CRF 模型的假设中，特征观测值不需要像 HMM 中要求的那样满足独立假设。除此以外，HMM 与 CRF 等模型还被使用在层次结构中，从而由简入繁、逐步识别出高层复杂动作^[66]。

2.4.2 未知异常行为识别

未知的异常行为是指那些事先并不能明确预知的各种突发可疑事件，这些行为在监控场合中虽然发生的机会很小，但却是特别需要重点关注的对象^[59,60]。由于现实中通常事先缺少这类行为的标注样本，人们往往借助一些无监督的学习算法来进行学习。这些方法的主要思想是：对已有的正常行为进行建模，如果发现新的行为与正常行为规律不同，则认为它属于异常行为。以下介绍几种主要的未知异常行为识别方法。

(1) 样本重构方法

这类方法比较直接，它主要是判断监控到的行为是否可以用已有的正常行为模板来重构。如果可以则代表测试行为不是异常，反之则为异常。例如 Boiman 等^[67]认为，如果一段目标视频中的行为不可以用已知视频库中的连续片段重构，那么该行为属于未知异常，反之则为正常行为。类似地，Cong 等^[68]提出了使用正常行为字典对测试样本进行稀疏重构的方法，并通过计算稀疏重构代价来衡量测试样本是否属于异常，而 Zhao 等^[69]样使用动态稀疏编码的方法实现对无监督的视频进行在线的异常行为识别。

(2) 聚类方法

聚类算法是一种很常见的无监督学习方法，它同样适用于未知异常行为识别。该类算法的主要思想是将正常行为的视频片断对应的描述特征进行聚类，得到一系列的簇中心。当一个新的行为片断抽取到的特征不属于任何簇，或者它对于所有主要的聚类中心都属于离群点时，则判定其为未知异常行为。除了经典的 k 均值 (k-means) 算法外，在其他一些工作中还用到 k 中心点 (k-medoids)、基于半径 (radius-based) 和基于蚁群 (ant-based) 等不同原理的聚类方法行为异常识别^[70-72]。

(3) 状态空间方法

产生式的状态空间模型同样可以很好地对视频中未知异常行为进行预估，其中 HMM 模型是使用最为广泛的产生式模型之一。正如前文提到的那样，HMM 模型非常有利于模拟时变的信息以及捕捉变量之间的潜在联系。当用在未知异常行为识别时，该类方法主要对不同的正常行为进行建模，然后对于新的行为序列，如果它们在已有的 HMM 模型上产生的概率越大，则表示其属于正常行为的概率越大，反之则代表它很有可能属于未知的异常行为^[73]。类似于 HMM 模型，Kim 和 Grauman^[74]将不同时间与空间上视频帧的区域特征作为马尔可夫随机场 (Markov Random Field, MRF) 的节点，然后用正常视频

上训练好 MRF 模型来计算产生测试视频行为的最大后验概率，从而判断每个视频区域上属于异常的程度，并设计了高效的更新机制来应对监控过程中的环境变化。

(4) 语义模型方法

还有一类未知异常行为检测的方法是基于语义模型的，如概率隐含语义分析模型 (probabilistic Latent Semantic Analysis, pLSA) 和隐含狄利克雷分布模型 (Latent Dirichlet Allocation, LDA) 等。这类语义模型本身也属于产生式模型，适用于无监督的数据。它们的基本思想是将底层的运动特征作为文档中的词语，共现的词语组成了主题，对应某一类行为；而共现的主题用来描述不同行为之间的联系。对于异常行为的判断原理与 HMM 模型类似，只要计算语义模型产生文档的概率大小即可。例如，Mehran 等^[75] 将一种转换成 BoW 的 social force 特征送入 LDA 模型中对人流中的异常行为进行识别，该种特征不仅考虑了人群的整体运动规律，还模拟了场景中人与人的互相作用，并假设人群相互作用的异常状态可以反映场景中异常行为的发生。Varadarajan 等^[76] 将行为的时间信息加入到了语义模型的建模环节中，以帮助异常行为的判断。

3 国内研究进展

3.1 监控视频编码

2000 年以前，我国一直采用视频编码国际标准。2000 年后，为了规避国际标准背后高额的专利费，工业和信息化部（原信息产业部）组织成立了数字音视频编解码技术标准工作组（简称 AVS 工作组），开始起草自主知识产权的国家标准《信息技术先进音视频编码》，并于 2006 年颁布为国家标准 GB/T 20090.2—2006^[3]，编码效率与同期国际标准 MPEG-4 AVC/H. 264 相当^[77,78]，对高清视频的压缩效率都能达到 150 倍。

2006 年的 AVS 国家标准颁布后，AVS 工作组开始着手面向行业应用对已颁布国标进行了定向扩展。从 2007 年开始，在 2006 年国标的基准档次（面向数字电视）基础上，相继扩展出加强档次（面向高清电影等应用）、伸展档次（面向视频监控等应用）和移动档次（面向手机流媒体等应用）三个部分。其中伸展档次（简称 AVS-S）是全球第一个针对视频监控应用制定的视频编码标准。

AVS-S 制定工作起始于 2007 年开始，需求分析是在公安部一所和视频监控行业多家企业共同参与下完成的，经过两年的努力，通过在基准档次的基础上增加适合监控视频特点的专用工具，于 2009 年完成了“伸展档次”（简称 AVS-S）。该标准针对视频监控全天候工作的特点，以监控现场的视频序列为测试基准，通过竞争方式选择、评估合适的视频编码技术制定而成。该标准不仅能够提高典型监控场景的编码效率，支持单色、彩色、红外序列编码，而且具有更强的抗误码特性和网络适应性，具有时域可伸缩性，

能够满足视频监控网络传输条件复杂的要求。该标准还提供了基于灵活条带和条带集的兴趣区域编码方法，能够支持图像区域标记、区域事件标记、摄像机标记等监控要求，并为感兴趣区域检测、对象分割和对象跟踪等智能应用和标准扩展预留了空间。

我国数字电视产业广泛使用 AVS 的重要原因是国外组织对采用国际标准的企业和运营商征收高额专利费，这个问题在视频监控行业并不明显，因此监控产业界转换到这样一个效率相当的新标准的动力不足。通过与视频监控行业的企业和应用单位的交流和调研，AVS 工作组判断，只有大幅度超越国际标准 H. 264，才能大幅度直接降低监控系统成本，新标准才有得到应用的可能。基于这个原因，2010 年 3 月，AVS 工作组启动了第二代视频监控标准（AVS-S2）的制定工作。AVS-S2 针对监控场景固定的特点，在传统基于块划分的混合编码框架的基础上，添加了基于背景帧的预测编码技术，形成了新的编码框架。与传统基于块划分的混合编码框架相区别，AVS-S2 的编码框架中包含新加入的背景建模单元、更新的基于背景帧的帧间运动补偿预测单元、背景帧缓存以及与背景建模和背景帧预测相关的控制逻辑，并在 2011 年底完成了标准起草工作。2012 年，面向立体电视和高清电视的 AVS + 标准制定完成，并被国家广电总局颁布为行业标准，AVS + 新增的一个重要工具是高级熵编码，这个工具也同样可以用于 AVS-S2。包含所有这些工具的新版 AVS 标准于 2012 年 10 月通过了 IEEE 标准委员会投票，于 2013 年 3 月获得批准，2013 年 6 月颁布为 IEEE 1857 标准。

AVS 标准的一个重要技术特色是针对应用需要制定简洁高效的标准方案和算法组合，2006 年颁布的 AVS 国家标准是针对数字电视需要而设计的，在变换、量化、熵编码、帧内预测、帧间预测和环路滤波等方面提出了一系列的新技术，在解码复杂度只有 H. 264 的 70%、编码复杂度只有 H. 264 的 30% 的情况下，获得了与 H. 264 相当的编码效率。

同 H. 265 再次提高编码复杂度的做法不同，AVS-S2 大幅度提高编码效率的主要“秘诀”是针对监控视频场景长期不变的特点，通过背景建模的方式去除了大量存在的“场景冗余”。背景预测代价冗余可以包含以下组成部分：视频暴露区预测编码冗余和块匹配过程中的前景干扰。所谓视频暴露区的预测编码冗余如图 3a 所示，当前图像的一些背景区域难以在最近参考图像或者长期参考的关键图像中找到相似的数据，但却可以通过建模产生背景图像，在背景图像中找到与这些背景区域更优预测参考。因此，研究在何种数据集上训练背景、使用何种背景建模方法建立背景、如何更新和编码背景、如何使用背景完成高效率的暴露区预测在监控视频编码领域中备受关注。块匹配过程中的前景干扰是因为现有视频编码标准（MPEG-2，H. 264，AVS，HEVC）都是以块为单位的，导致监控视频中的前景、背景像素混合数据的编码效率较低。一个典型的例子如图 3b 所示，前背景混合的数据块 A 与它的匹配参考 A' 中的背景部分不具有相似性；数据块 B 与它的匹配参考 B' 的前景部分不具有相似性。虽然 H. 264 和 HEVC 拥有较小的预测块大小，但它们在显著增加编码复杂度的同时，这一问题并没有得到根本解决。因而，在块匹配的编码框架基础上，如何提高前背景混合数据块的预测编码效率，也是急需解决的重要问题。



图 3 监控视频进行预测编码时的两个问题

针对暴露区的预测效率问题，文献[79]提出了基于背景建模的视频编码算法框架，如图4所示。首先，训练集图像经过背景建模、背景编码、背景解码后，生成高质量的重建背景图像；然后，编码器使用该重建背景完成对每幅图像进行背景预测以降低背景部分的编码代价，进而得到更优的编码效率；最后，背景图像码流和视频序列码流都传输给解码过程用于解码。在该编码框架下，文献[80]提出了适用于视频编码的背景模型，为了进一步提高暴露区域的编码效率，文献[81]针对HEVC的低延迟编码配置，提出了

基于背景模型的视频图像位率分辨率编码优化算法。

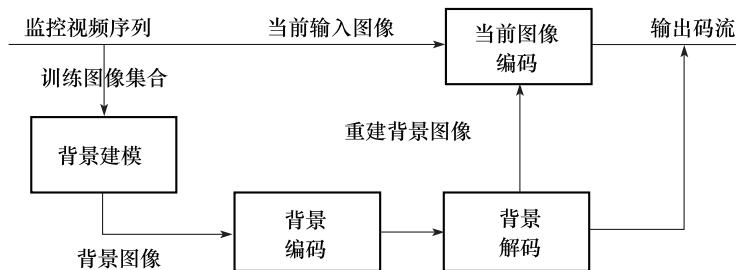


图 4 基于背景建模的监控视频编码框架

为了解决背景差分编码在前景较多的序列中性能增益较少的问题，文献[82]提出了基于宏块分类和背景模型的自适应运动补偿算法。在该算法中，可选择的运动补偿包括：BRMC，使用高质量编码的使用原始图像训练生成的背景作为长期参考图像；BDMC，使用当前宏块的参考数据和这些参考数据对应的背景像素的差来运动补偿预测当前宏块和其背景的差分结果，如同文献[17]所述的背景差分预测方式；SRMC，使用最近解码的参考图像来预测当前宏块。

集成上述所有技术的 AVS 监控视频编码标准已经作为 AVS 视频标准独具特色的一个档次于 2013 年颁布为 IEEE 1857 国际标准。IEEE AVS 标准中的监控档次是第一代 AVS 技术的集大成者，是全球第一个面向视频监控的国际标准。

以 10 个典型监控视频作为测试序列，将 IEEE 1857 监控档次和国际标准 H.264 的高级档（High Profile）、AVS 国家标准基准档（GB/T 20090.2-2006）和 2012 年发布的 AVS 广播档（AVS+）进行对比，对比软件均采用这些标准最新版本的参考软件。实验表明，在压缩这些监控视频序列时，IEEE 1857 监控档次与其他三个标准档次相比，平均码率节省都超过了 50%，即编码效率是它们的两倍。

3.2 目标检测与跟踪

国内许多高校和研究机构在基于视频监控的运动目标检测与跟踪方面已投入了大量研究精力。中科院自动化所模式识别国家重点实验室成立了智能视频监控研究组，研究交通场景监控、人脸检测与跟踪、多摄像机联合跟踪和异常行为检测等，该团队开发的全天候实时智能视频监控技术已应用到实际中，促进了安防产业的发展，2011 年“面向安全监控的视频内容理解技术与应用”课题获得国家科技进步奖二等奖。北京航空航天大学数字媒体北京市重点实验室在视频图像的分析处理方面开展了多年的研究工作，提出了基于状态感知的视频运动对象分析方法，实现了 3×3 像素以上运动目标检测，平均正确率达到 99%，误报率低于 1%，在相似物干扰、光照变化、转弯（跳变）等复杂情况下仍然能正确跟踪目标，相关研究成果已用于我军装备，并以该技术为基础之一获得了 2011 年国家技术发明二等奖。西安电子科技大学智能感知与图像理解实验室针对高分辨 SAR 图像，引入核学习机并完成了多脊波网络的构造与逼近性的研究，提出了自适应

脊波网络和自适应学习算法、基于决策树的支撑矢量机多类模式识别算法，对 SAR 图像中的桥梁、港口和机场三类目标的平均识别率分别达到了 92.00%、94.29% 和 93.33%。此外，在国家自然科学基金的资助下，2010 年北京大学的田永鸿等开展了基于多摄像头协同的运动目标检测跟踪和异常行为分析的研究；2012 年，上海交通大学的杨杰等开展了多视频摄像头组网协同下目标检测分析的关键技术研究，华中科技大学的桑农等开展了基于遮挡分层模型的遮挡目标跟踪技术研究，云南大学的周浩等开展了基于行为模式分析的戒毒所智能视频监控关键技术研究；2013 年，电子科技大学的叶茂等开展了基于特征学习的领域自适应目标检测方法研究，北京航空航天大学的张兆翔等开展了基于多任务学习的跨视角智能视频分析方法研究等。

与此同时，在运动目标检测跟踪、视频内容理解、语义挖掘等课题研究的支撑下，相关技术发展也有了较大的进步。例如，针对低码率视频中的快速目标检测和跟踪问题，清华大学的艾海舟教授等人^[83]提出了融合多尺度多生命周期观察者模型的概率粒子滤波技术，取得了明显优于传统方法的检测精度，该方法获得 2007 年 CVPR 会议最优学生论文奖。中科院自动化所模式识别国家重点实验室谭铁牛研究员和黄凯奇博士提出异构数据融合机制，所带领的智能视频监控研究团队在 2010 年和 2011 年国际著名视觉识别竞赛（PASCAL-VOC）中击败卡内基梅隆大学、斯坦福大学以及微软等国际顶级研究团队，蝉联目标检测冠军和图像目标分类亚军。复旦大学构建了一种利用目标局部中层表示的视觉特征，并将其应用于图像场景识别任务^[84]，提出基于轨迹和运动参考点的动作识别，利用物体（人体）以及背景之间的相互运动关系提高识别的准确率^[85]，在 2012 年 MediaEval（欧洲最知名的多媒体分析与检索的评测平台）视频暴力事件检测任务中战胜英国帝国理工、法国 INRIA、日本 NII 等团队，取得了最优的精度。

在国家大力支持下，国内也成功举办了大量智能视频监控相关的国际会议。2002 年、2003 年和 2011 年召开的全国智能视觉监控学术会议，围绕智能视觉监控领域中的运动检测与跟踪、物体识别与分类、行为理解、多传感器融合、硬件系统集成等主题展开了交流。由中国计算机学会多媒体技术专业委员会、中国图象图形学学会多媒体专业委员会、中国计算机学会普适计算专业委员会、中国自动化学会计算机图形学和人机交互专委会、ACM SIGCHI 中国分会共同发起和主办的全国和谐人机环境联合学术会议（HHME）截止 2014 年已经连续召开了 9 届，其中的多媒体分会为计算机视觉、视频图像分析处理的研究者提供一个交流创新思想、展示研究成果的平台。

在获得上述成果的同时，我们也意识到虽然国内对运动目标检测与跟踪的研发已有不少，在一些技术点方面已经具有国际先进水平，但是由于国内起步较晚，在系统的实用性、推广应用方面还有待提高。

3.3 监控视频增强

3.3.1 去雾

随着智能监控系统在国内的大范围普及，低质视频增强也是一个研究热点。国内

关于图像去雾算法的研究较晚于国外，但是随着国内对图像去雾领域研究的不断重视、国际合作越来越紧密，国内的研究者在国际重要学术期刊以及国际顶级学术会议上发表了越来越多的高水平论文，如香港中文大学的何恺明^[86]、中科院自动化研究所的孟高峰^[87]等。尤其值得注意的是何恺明提出的基于暗原色先验知识的单幅图像去雾方法^[66]获得了计算机视觉与模式识别顶级会议 IEEE CVPR 2009 的最佳论文奖，这是 CVPR 创立以来首次由中国人获得这个奖项。

3.3.2 去夜

在图像夜色去除方面，很多国内高校和科研院所都陆续展开研究工作，比如中国科学院自动化研究所、北京邮电大学和西北工业大学等。稀疏表达是近年来计算机视觉和机器学习领域中的一大研究热点，已被成功地应用于图像去噪等场合。文献[88, 89]对前景目标进行提取，根据白天与夜晚背景图像进行融合，该方法对传统的像素级别融合方法灵活度不够的缺点进行了改进。为了对单张图像进行去夜，文献[90]提出了一种融合了颜色模型与稀疏编码的算法框架，对图像前景目标的噪声有较好的鲁棒性，能够得到边缘增强的复原效果。

3.3.3 去雨雪

国内相关研究人员对国际上已有的雨雪去除的方法进行了系统的分析和研究，并做出了一些有效的改进工作。文献[91]改进了逐像素处理方法，将彩色图像的像素点的 RGB 三个通道分别处理，采用投影寻踪模型进行分类，检测雨雪像素并去除。文献[92]提出一种改进 snake 模型，对视频图像中雨雪颗粒的轮廓进行识别，从而进行雨雪的检测和去除。

3.3.4 去模糊和超分辨率

在超分辨率和去模糊方面，国防科技大学、电子科技大学、浙江大学和中科院等多个高等院校和科研所进行了相关研究。国内权威期刊如计算机研究与发展、自动化学报等也发表了相关研究成果。浙江大学张小红^[93]针对物体平面运动引起的物体模糊，提出了一种适用于双图像的去平面运动模糊方法。浙江大学唐磊^[94]通过研究低分辨率图像退化矩阵的联系，提出了联合运动估计与超分辨率重建算法。

3.4 视频动作与异常行为识别

我国较早开展相关领域研究的是中国科学院自动化研究所模式识别国家重点实验室等单位，现在国内基于视频动作分析的研究正蓬勃发展，很多高校已经有许多专门从事智能视频监控技术的课题组。

近年来，国内学者在国际权威杂志与顶级会议上发表了许多出色的工作。这些工作的内容涵盖了视频中动作监控的各个方面，如特征的设计、分类模型的改进、上下文信息的利用和算法效率提升等。举例来说，近期在 Wang 等^[95]使用密集轨迹与基于运动边界直方图的特征描述对不同数据集上的动作识别进行了评测，取得了理想效果，影响很大。Jiang 等^[96]提出了一种基于密集轨迹的人体动作识别特征，该方法将一小段时间内所有的轨迹根

据运动大小和方向聚类获取全局参考点，找出主要的运动方向和大小，作为背景的运动，然后统计轨迹点之间的相互运动的大小和方向获取局部参考点，最终所有轨迹根据背景运动来修正，以获取更佳的动作描述特征。文献[97]针对人与人行为之间的相互作用的问题，采用层次结构的 CRF 模型来模拟人的身体部位之间的运动关系以及人与人之间的关系。在 Ouyang 等^[98]使用了混合可变形部件模型（Deformable Part-based Models, DPM）对画面中邻近的多个动作进行了识别。在 Sun 等^[99]采用 Kinect 设备获取了带有深度特征的原始视频，并使用 latent SVM 模型对视频中的手势动作进行了识别。Lu 等^[100]则是专注于提升算法的实时运行效率，他们对人体动作的稀疏重构环节进行了算法改进，在不牺牲识别性能的前提下，将运行速度提升到了 150 FPS 的程度。Jiang 等^[101]系统地总结了视频中高层事件检测相关的各个环节的技术思路与要点，具有一定参考价值。

此外，在近年国内的重要刊物上也有一些最新的相关工作陆续发表，这些工作同样涵盖不同的主题，包括但不限于特征表示^[102,103]、分类策略^[104,105]以及不同的应用环境^[106]等方面。

4 国内外研究进展比较

4.1 监控视频编码

基于背景建模的编码方法实质上是消除常规标准没能消除的“场景冗余”，因此同样用于提高其他视频编码标准的效率。我们将这套方法增强国际标准 HEVC (H. 265)，同样用上述十个监控视频序列和 HEVC 参考软件进行对比，实验表明能将 HEVC 的码率平均再降低 44.78%，而且复杂度降低 46.53%，即用约一半的复杂度实现了编码效率的翻番，压缩效率达到现行国际标准 H. 264 的近四倍。

基于背景建模的监控视频编码技术也已经被第二代 AVS 标准 AVS2 采纳。AVS2 研究工作开始于 2009 年，2012 年 3 月正式发出提案征集书，经过两年的努力，2014 年 4 月完成标准起草。基于背景建模的编码技术是 AVS2 的主要组成部分，使得 AVS2 在编码背景固定的监控视频、教学视频、法庭视频、风景视频等场景类视频时，比同期国际标准 HEVC/H. 265 高一倍。对于视频会议乃至常规电视视频，在背景稳定的情况下，编码效率也明显比 HEVC 高。因此，AVS2 决定把这项技术作为基本工具，即所有 AVS2 解码器和编码器都支持这个功能，从而使得 AVS2 明显领先同期国际标准，达到监控视频编码效率的最高水平。

4.2 目标检测与跟踪

在计算机视觉领域中对视频内容的分析，主要以视觉中的运动目标为主，即以运动

目标为中心来检测其存在位置、动作变化，以及与周围环境的关联和互动。因此，运动目标检测与跟踪具有广阔的应用前景。特别是美国 911、俄罗斯地铁恐怖袭击、英国电车恐怖袭击事件以来，人们对安全的需求越来越强烈，利用各种传感器拍摄视频进行监控已开始广泛用于各领域。国外很早就对这一领域极为关注，早在 1997 年，美国国防高级研究计划署（DARPA）就资助卡内基梅隆大学和萨尔诺夫戴维研究中心等著名高校和研究机构，联合研制了视频监控系统 VSAM，该系统针对旋转云台和机载摄像机视频，可以检测和跟踪车辆目标；2010 年，DARPA 启动一项红外视频目标检测与跟踪系统 ARGUS-IR，期望实现黑夜条件下对大范围的战场环境或无人区域的侦察与监视。此外，法国 INRIA、美国马里兰大学、英国伦敦大学、以色列 IOImage 公司等在视频目标检测跟踪领域也开展了一系列创新研究工作，并取得了一些具有重要影响的成果。在国内，中科院自动化所、中科院计算所、清华大学、北京大学、北京航空航天大学、西安交通大学、复旦大学、哈尔滨工业大学等一些高校和研究机构近年来也在视频目标检测跟踪领域开展了卓有成效的研究工作。

比较而言，国外的产业化发展更好，例如 IOImage 作为全球最出色的视频分析设备厂商，开发的智能视频监控系统具有 PTZ 自动跟踪、入侵检测、警戒线越界、跨越围栏、丢包等行为的识别，而国内现有系统基本以此为标准进行开发设计。从技术方面来看，近年来，随着国际合作越来越紧密，国内的研究者在国际重要学术期刊以及国际顶级学术会议上发表了越来越多的高水平论文，已与国外的研究处于同步阶段。在一些运动目标检测跟踪算法的测评活动中，也有许多国内研究人员的参与，如 2014 年的 IEEE Workshop on Change Detection。

总的来说，视频运动目标检测跟踪是视频图像内容理解中一个实用性较高的功能，已形成较好的产业化。但是受复杂场景的影响，以及算法实时性的要求，要做到准确性高、适应性好的视频运动目标检测跟踪仍存在许多难以解决的问题。目前，视频运动目标检测跟踪的研究主要集中在有效特征提取、分类器训练等方面，近年来，机器学习等方法的引入，使得基于多实例学习、稀疏表示的运动目标检测跟踪成为研究的热点。此外，检测跟踪算法的评估也是一个需要关注的问题，虽然 Goyette^[107] 在 2012 年提出了一个用于运动目标检测评估的数据库，包括了室内、室外场景共 31 段视频，涵盖了船只、小轿车、卡车、行人等多样情况，Wu^[108] 在 2013 年提出了包括 50 段视频序列的运动目标跟踪评估数据库，涵盖光照变化、目标尺度变化、遮挡、形变、运动突变、旋转、消失、低分辨率等复杂情况，但是发展被广泛接受的、实际可用的、大规模的视频数据还是一个努力的方向。相对来说，国内在这方面的研究较少，缺乏系统性。

4.3 监控视频增强

低质视频增强是智能视频监控领域重要的研究方向，其意义在于保证智能视频监控系统不仅能在一般环境条件下工作，还可以在各种复杂环境下正常工作，具有十分重要的应用价值和广阔的应用前景。与传统的图像或者视频增强方法不同，监控视频

质量增强概念的外延更广，它主要包括了去雾、去夜、去雨雪、去模糊、超分辨率增强等多方面的内容，这几个方面的研究具有相似的应用背景，但是如上文所述，往往在方法上有很大的差异性，因此在实际处理过程中具有领域交叉性和算法复杂性的特点。

在国内，从事低质视频增强的单位主要有浙江大学、清华大学、北京大学、北京航空航天大学、北京邮电大学、中国科学院计算技术研究所、中国科学院自动化研究所、微软亚洲研究院、香港中文大学等。相关课题得到了国家自然科学基金等的资助。近年来，随着国际合作越来越紧密，国内的研究者在国际重要学术期刊以及国际顶级学术会议上发表了越来越多的高水平论文。在国外，率先关注图像增强这一领域的研究单位主要有：麻省理工学院、美国哥伦比亚大学、美国伊利诺伊大学、美国卡内基梅隆大学、美国哥伦比亚大学以及牛津大学等。相比较而言，国外研究人员更加关注视频增强的理论研究，相关工作在国际学术刊物（International Journal of Computer Vision、IEEE Trans on Pattern Analysis and Machine Intelligence、IEEE Trans on Image Processing）和重要国际会议（International Conference on Computer Vision、IEEE Conference on Computer Vision and Pattern Recognition）上发表量逐年增加。相比较而言，国外研究内容更偏向理论，而国内的很多研究工作建立在应用导向之上，在借鉴国外相对成熟的理论体系和技术应用体系的条件下，国内的增强技术和应用也有了很大的发展。

目前，图像去雾主要采用的方法可以分为两类：基于图像处理的增强方法和基于物理模型的复原方法，国内与国外在这个领域的研究进展处于同步状态。基于图像处理的方法主要是从图像处理角度入手，通过统计有雾图像的时域和频域特性，选取适当的图像增强算法来提高图像的清晰度。基于物理模型的方法主要是从有雾图像的物理成因角度入手，通过简化或改进大气散射和视觉成像等物理模型进行图像去雾处理。在去夜色方面，由于没有统一的数学模型，因此国内外在这个领域的研究呈现百花齐放的状态，研究方法大多以应用为导向，探索在不同实际情况下的低照度视频增强方法。在雨雪等天气去除方面，由于雨雪等颗粒的检测是天气去除问题的关键，研究人员提出了很多雨雪颗粒检测方法。去模糊和超分辨率增强一直是计算机视觉领域的热点和难点问题，相对来讲，国内研究者虽然取得了很多研究成果，但与国外研究水平尚有一段距离，如何去寻求新的图像观测模型、运动估计模型以及图像先验分布模型对于超分辨率重建算法来说至关重要。总体而言，现有的各种去雾、去夜色、去雨雪以及去模糊算法均是对某类图像清晰化效果较好，而对其他类则相对较差。因此，应探索具有较好普适性的视频增强算法。

4.4 视频动作与异常行为识别

相比之下，国内对于视频监控中的智能分析研究起步较晚。近年来涌现出一批优秀的科研成果，发表在领域内顶级的期刊与会议上，但是，从总体上，研究水平与欧美等发达国家相比仍存在一定差距。这尤其体现在创新能力上，许多真正的开创性工作还是

集中在国外的一些著名研究机构的成果中。从研究内容上说，国内和国外的研究者都开始了对于相对复杂环境下的复杂动作进行识别的尝试，并且在策略上尤其重视对于上下文信息的利用，国内外差距不大。但是，目前方法对于复杂的识别效果都还没有达到大规模实用的要求，在算法层面上还有很大的提升空间。

国外著名的高校与研究组较多，特别是在欧美国家，比如牛津大学的 VGG 研究组、美国中佛罗里达大学的 CRCV 研究组，以及法国国家信息与自动化研究所（INRIA）的 LEAR 研究组等均长期活跃在科研一线，他们的工作对整个领域的技术发展有着重要的引领与推动作用。

国外高校与公司的合作也较多，很多公司如 Sarnoff、BBN、ObjectVision 等都与很多高校合作。同时，国外专注于该领域深入研究的公司也比国内要多，如 IBM 公司就专门设有对应的研究部门，微软公司围绕 Kinect 产品的科研成果很大程度推动了大量基于视觉的动作识别技术的发展与应用。相比之下，国内高校与公司之间的深入合作仍然较少，国内致力于监控领域基础算法研究的公司相比国外就更少了。

5 发展趋势与展望

5.1 监控视频编码

在视频编码 60 多年的研究历程中，形成了熵编码、变换编码和预测编码三类方法。过去 30 多年，基于这三类方法制定了以 MPEG-2（1994）、AVC/H. 264（2003）、HEVC/H. 265（2013）为代表的三代标准，其中后一代标准比十年前的前一代标准编码效率提高一倍。分析三类编码方法对三代标准效率翻番的贡献可以看出，贡献最大的是预测编码：在 MPEG-2 中用来去除时间冗余（帧内预测），在 AVC/H. 264 中进而用于去除空间冗余（帧间预测），在 HEVC/H. 264 和可变尺寸块及更复杂的预测模式相结合，再次对效率提升做出主要贡献。从国内外的研究状况来看，未来的监控视频编码的主要发展趋势表现在以下几个方面。

1) 基于背景建模的监控视频编码方法。该方法拓展了预测编码，实现了（背景区域的）模型预测和（前景的）差分预测，为预测编码这棵“老树”增加了一个生机勃勃的“新枝”，“单枪匹马”实现了监控视频压缩效率再次翻番，这一思路可以继续扩展。

2) 基于场景建模的监控视频编码方法。以监控视频为代表的场景视频，其背景和前景中存在大量不变的或规律性变化的部分，因此可以采用机器学习的方法得到日益完善的背景模型库和前景对象库，即场景模型。场景模型是一个比背景图像更丰富的动态模型库，它会随着时间的延伸而日益完善，新的规律性对象将会入库，而当前视频中与场景模型不符合的部分才需要进行编码、传送和存储，这样就可以进一步提高视频压缩效

率。简而言之，未来的视频码流变成了一个共性内容的模型库或知识库（相同信息只存一次）和一个只记录画面变化信息的压缩码流，从而明显提高预测性能和压缩效率。

3) 云视频编码。除了场景视频编码器自行生成和完善模型库外，利用互联网上广泛存在的图像和视频构造具有通用性的模型库同样可以提高预测性能，我们称之为“云视频编码”，已经得到国家自然科学基金重大项目支持，也是下一代 AVS 标准努力突破的重要方向。

5.2 目标跟踪与检测

现有视频运动目标检测跟踪方法主要对中低层特征进行处理，易受场景噪声、场景和目标状态多变、目标类型多样等影响，存在准确性低、适应性差等问题，在检测时往往难以在准确检测到慢速目标或运动小目标的同时排除树枝叶晃动等背景扰动，以及雾气、沙尘等噪声干扰；在跟踪时往往适应简单背景的目标，在复杂背景以及存在多运动目标、目标被遮挡、目标作无规律运动以及形变时难以进行连续准确、稳健的跟踪，系统的实用性是最大的问题。从国内外的研究状况来看，运动目标检测跟踪的不足与发展趋势主要表现在以下几个方面：

1) 突破中低层特征的局限。中低层特征不能有效表示图像中的丰富信息，突破像素级处理和中低层特征分析方法的局限，研究结合场景信息和目标状态的视频运动分析方法，提高算法的实用性。

2) 综合利用各种互补信息。研究时域、空域、频域信息，以及不同尺度空间特征信息的结合，提高检测的准确性。

3) 结合目标全局与局部特征。特征对目标的描述不够准确，特征对遮挡、相似物干扰等状态的辨别能力较弱，不能很好适应复杂场景。需要构建特征提取新算子，特别是不满足运动连续性假设时的有效特征提取算子。

4) 消除特征时间的干扰。特征作用会随着目标状态的变化而变化，盲目集成多个特征有时反而会降低跟踪系统的稳健性，尤其是当其中一部分特征受噪声、相似物干扰导致有效性降低时。跟踪算法应研究根据这些变化动态选择有效特征的机制。

5) 准确的相似性计算方法。现有方法进行特征匹配时采用传统的相似性度量函数计算，但受遮挡、噪声等影响，相似度计算结果并不可信，如何利用准确的相似性计算方法确定特征融合权值是亟需解决的问题。

5.3 监控视频增强

在去雾方面，未来的研究重点与趋势主要在以下两个方面：

1) 提高算法的普适性。对于已有的图像去雾方法而言，现有方法各有优缺点，针对的应用的场景也千差万别，但各种去雾算法均是对某类图像清晰化效果较好，而对其他类则相对较差。基于物理模型的图像复原方法尽管已取得较大进展，但是由于景物退化

与场景深度呈非线性关系，由此带来的一个最大问题是很难保证所建立的景物退化模型的普适性。因此，探索具有较好普适性的图像去雾算法和研究基于更完备的物理模型的去雾算法在未来一段时间内都将是一个具有挑战性的课题。

2) 提高算法的处理速度。由于图像去雾任务本身通常含有大量复杂的数据处理算法，例如大型矩阵的分解，大规模方程组的求解以及众多的非线性优化问题。这些复杂运算往往需要较长的处理时间，但对于实际应用来说，算法的实时性是至关重要的。由于现有算法在某些对实时性要求高的场合无法达到预期的执行速度，所以一方面要重点研究图像去雾处理的简易、快速优化算法，另一方面要研究如何利用并行化技术，实现多处理机快速处理的并行算法。此外，利用可编程图像硬件加速图像去雾算法也是未来研究的一个热点。

在雨雪去除方面。尽管视频中雨雪去除已进行了大量研究，雨雪的图像域和频率域模型已基本建立，众多科研人员也提出了许多有效的方法，但已有的工作仍仅限于雨雪量较小，背景较为简单的场景，方法的测试和评估都是建立在哥伦比亚大学视觉实验室(Computer vision laboratory of Columbia University, CAVE)的数据集上，缺乏普遍性和说服力。各种方法都还未考虑算法的时间复杂度，目前各算法的处理效率还很低，难以用于现实场景及实时视频处理当中。因此，该领域还应在以下几方面进一步研究：

1) 雨雪建模。雨雪建模是整个研究内容的理论基础，已有许多研究人员对雨雪的物理模型和运动模型进行了分析和建模，但已有模型都是建立在环境光线稳定、风速变化较小、雨雪直线运动的情况下，仍无法描述环境光照变化及风速对雨雪运动的影响。这一点是需进一步研究的方面。

2) 雨雪识别。雨雪识别是雨雪去除的前提，研究人员已经对雨雪颗粒的物理模型和运动模型进行了建模，但已有方法仅在雨雪较小、背景纹理较简单、背景物体无运动或运动较慢的情况下有效，如何准确识别雨雪颗粒，减少背景运动物体的干扰，仍是需要进一步研究的重点。

3) 鲁棒性和实时性的提高。在已有模型下，已有的方法对环境光线变化、风速变化、复杂背景、运动物体等条件的鲁棒性较差，很难应用于普遍场景和实际应用中。此外，已有的算法时间复杂度较高，无法应用于实时处理中，如实时户外监控中。如何提高算法与系统的鲁棒性与实时性，是视频中雨雪去除问题的重点和难点。

在去夜色和去模糊方面，未来的研究重点与趋势主要在以下几点：

1) 算法的实时性。监控视频的一个重要需求是监控的实时性。好的夜色去除算法应该满足复杂度低、能够实现在线的监控场景照度增强，而不依赖于过多的离线训练过程的特点。

2) 特定区域增强。在进行整体图像增强的同时，应当对监控中最感兴趣的诸如前景目标、人脸等关键区域进行增强或者复原。

3) 与去模糊技术和超分辨率增强方法相结合。经过去夜后的视频可能出现局部信息缺失和模糊的现象，未来应将各种去模糊、图像增强、超分辨率技术与去夜技术相结合，使得低质视频处理更加高效。

5.4 视频动作与异常行为识别

现有的智能动作分析与异常行为识别技术虽然得到了不断发展，算法的性能也在不断提高，但是从实用角度，除了简单的特定或可控场景外，拥有动作识别和异常行为识别能力的视频监控系统还没有太多成熟的应用案例。技术上，对于复杂场景与复杂动作的识别精度还远没有达到实用化水平。该领域未来的发展趋势主要体现在以下几个方面：

- 1) 新特征的设计与使用。需要创新的不仅在于更好的识别算法，其中特征设计或学习非常重要。一些最新的工作均把创新点放在如何设计出新颖的特征上，因为在使用机器学习的算法做最终的分类时，特征的好坏将决定性能的上限。鉴于此，设计或自动学习出更好的特征描述方法仍将是未来极为重要的技术创新之处。
- 2) 深度信息的广泛使用。在数据采集层面，随着 Kinect 之类的 RGB-D 采集设备的发展，有深度信息的 RGB-D 在许多传统的机器视觉领域都有了成功应用。在未来一段时间，随着 RGB-D 的采集设备的进一步普及，在视频监控系统中更好地利用 RGB-D 数据将是一个非常值得关注的研究课题。
- 3) 深度学习技术的使用。近年来深度学习技术的发展得到了全世界的普遍关注。深度学习方法已经在语音识别和图像分类等领域取得了突出的成绩，并已经有一些研究者将该技术应用到视频内容理解中来。但是在监控视频中的应用以及异常检测中目前还没有太多这方面的工作，可以期待将深度表示的特征与动作识别的算法相结合来获取更好的识别效果，其中可能的难点将是如何更有效的在特征的提取中更合理的融合动作序列的时间信息，同时又要将神经网络的复杂度控制在可接受的范围内。
- 4) 重视特定上下文信息的使用。鉴于开放环境下的复杂动作识别方法一直以来都是研究的难点问题，许多最新的工作成果都在讨论如何用上下文信息（如场景、目标的检测与识别）来帮助相对复杂的动作识别。对于不同的应用场景与不同的类别，可以期待在未来的工作中发掘出更多的特定上下文信息来帮助提升识别效果。
- 5) 监控视频数据标注。随着大数据时代的到来，越来越多的研究者更倾向于使用大规模的数据来训练分类模型，并在大量测试数据上验证算法的有效性。特别是对一些基于统计模型或深度网络结构的方法，在拥有更多的数据的情况下，总是可以期待更好的训练效果。现有的监控视频训练数据集，无论从数量还是种类上，都还是比较有限的，远没有互联网数据或像图像数据那样多，很大原因在于高质量的视频数据的标注成本相对而言非常昂贵。为了推动该领域更好的发展，需要更多的研究团体积累与发布更多、更全面的已经标注的视频数据供学界使用，为识别算法研究提供有力的数据保障。

6 结束语

目前，我国最新制定的国家标准 AVS2 在对监控视频的编码效率上比最新国际标准

H.265/HEVC 高出一倍，标志着我国的视频编码技术和标准在视频监控领域已经实现跨越。这为大幅度降低监控视频的传输和存储成本创造了巨大的技术和产业机遇。在运动目标检测和识别上，目前的方法易受场景噪声、场景和目标状态多变、目标类型多样等影响，存在准确性低、适应性差等问题，并且方法的实用性面临的最大问题。在监控视频质量增强上，目前的研究思路是对视频图像进行后续处理，而如何快速实时地对图像进行快速校正，从而实现监控视频的实时处理需要进一步研究。海量监控视频中动作与异常行为分析是一个非常复杂的研究问题，涉及计算机视觉、机器学习等多个领域的技术。虽然近年来该领域的技术发展突飞猛进，但是离真正实用的自动识别系统还有很长的路要走。随着大数据时代的到来，智能视频监控的需求将日益迫切，面对众多挑战的同时，该研究领域将迎来前所未有的重大机遇，将会有越来越多的研究者关注该领域，也必将产生越来越多可以使用的研究成果。

参考文献

- [1] ITU-T. Advanced video coding for generic audiovisual services. ITU-T Rec. H.264, 2004.
- [2] ISO/IEC 14496-10 Information technology—Generic coding of audio- visual objects - part 10: Advanced video coding. 2004.
- [3] GB/T 20090.2-2006. 信息技术先进音视频编码第 2 部分：视频. 2006.
- [4] Bossen F, Bross B, Sühring K, et al. HEVC complexity and implementation analysis. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1684-1695.
- [5] Musmann H G, Hotter M, Ostermann J. Object- oriented analysis- synthesis of moving images. Image Communication, 1989, 1(2): 117-138.
- [6] Wang J Y, Adelson E H. Representing moving images with layers, IEEE Transactions on Image Process, 1994, 3(5): 625-638.
- [7] Chai D, Ngan K. Foreground/background video coding scheme. In: Proceedings of IEEE International Symposium Circuits Systems, 1997, 1448-1451.
- [8] Martins I, Corte R L A video coder using 3-D model based background for video surveillance applications [C]. In: Proceedings of IEEE International Conference on Image Process, 1998, 919-923.
- [9] Richardson I E G . H.264 and MPEG- 4 video compression: video coding for next generation multimedia. England: John Wiley & Sons Ltd, Chichester, 2003.
- [10] Francois E, Vial J F, Chupeau B. Coding algorithm with region- based motion compensation. IEEE Transactions on Circuits Syst. Video Technol, 1997, 7(1): 97-108.
- [11] Vetro A, Haga T, Sumi K, et al. Object-based coding for long- term archive of surveillance video [C]. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2003, 417-420.
- [12] Babu R V, Makur A. Object-based surveillance video compression using foreground motion compensation [C]. In: Proceedings of International Conference on Control, Automatic, Robotics and Vision, 2006, 1-6.
- [13] Hakeem A, Shafique K, Shah M. An object-based video coding framework for video sequences obtained from static cameras [C]. In: Proceedings of ACM International Conference on Multimedia, 2005,

- 608-617.
- [14] Morde A, Ma X, Guler S. Learning a background model for change detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2012, 15-20.
- [15] Dong Z, Javed O, Shah M. Video Object Segmentation through Spatially Accurate and Temporally Dense Extraction of Primary Object Regions[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2013, 628-635.
- [16] Zhou X, Yang C, Yu W. Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013, 35(3) : 597-610.
- [17] Wang D, Lu H, Yang M. Least Soft-threshold Squares Tracking[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013, 2371-2378.
- [18] Zhang K, Zhang L, Yang M H Real-time compressive tracking[M]. In Proceedings of Computer Vision-ECCV, Springer Berlin Heidelberg, 2012: 864-877.
- [19] Liu B, Huang J, Yang L, et al. Robust Tracking Using Local Sparse Appearance Model and K-Selection [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011: 1313-1320.
- [20] Barnich O, Droogenbroeck M V. ViBe: A universal background subtraction algorithm for video sequences. IEEE Transactions on Image Processing, 2011, 20(6) : 1709-1724.
- [21] Leichter I. Mean Shift Tracker with Cross-Bin Metrics. IEEE Transaction Pattern Analysis and Machine Intelligence, 2012, 34(4) : 695-706.
- [22] Babenko B, Yang M H, Belongie S. Robust Object Tracking with Online Multiple Instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8) : 1619-1632.
- [23] Toyama K, Krumm J, Brumitt B, et al. Wallflower: Principles and practice of background maintenance[C]. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, 1: 255-261.
- [24] Zhao C, Wang X, Cham W K. Background subtraction via robust dictionary learning[J]. EURASIP Journal on Image and Video Processing, 2011(2) : 1-12.
- [25] Lu C, Shi J, Jia J. Online Robust Dictionary Learning[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013: 415-422.
- [26] Baker S, Scharstein D, Lewis J P. A database and evaluation methodology for optical flow [J]. International Journal of Computer Vision, 2011, 92(1) : 1-31.
- [27] Kurmi U S, Srivastava H S, Agrawal D, et al. Performance Evaluation of RGB Skin Color Segmentation Based Face Detection Technique[J]. International Journal of Engineering Universe for Scientific Research and Management, 2014; 6(2) : 1-6.
- [28] Rougier C, Meunier J, St-Arnaud A, et al. 3D head tracking for fall detection using a single calibrated camera. Image and Vision Computing, 2013, 31(3) : 246-254.
- [29] Aparna A, Ripul G, Satish K, et al. Moving target detection in thermal infrared imagery using spatiotemporal information. 2013, 30(8) : 1492-1501.
- [30] Cheng M M, Zhang G X, Mitra N J, et al. Global contrast based salient region detection [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2011: 409-416.
- [31] Li J, Levine M D, An X, et al. Visual saliency based on scale-space analysis in the frequency domain. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(4) : 996-1010.
- [32] Lowe D. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer

- Vision, 2004, 60(2) : 91-110.
- [33] Liu C, Yuen J, Torralba A. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(5) : 978-994.
- [34] Rublee E, Rabaud V, Konolige K, et al. ORB: an efficient alternative to SIFT or SURF [C]. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. 2011: 2564-2571.
- [35] Kinjal A J, Darshak G T. A Survey on Moving Object Detection and Tracking in Video Surveillance System [J]. *International Journal of Soft Computing and Engineering (IJSCE)*, 2012, 2(3) : 44-48.
- [36] Mei X, Ling H, Wu Y, et al. Minimum error bounded efficient l1 tracker with occlusion detection. In: *Proceedings of CVPR*, 2011: 1257-1264.
- [37] Avidan S Ensemble tracking. In: *Proceedings of CVPR*, 2005, 494-501.
- [38] Zhang K, Song H. Real- time visual tracking via online weighted multiple instance learning. *Pattern Recognition*, 2013, 46(1) : 397-411.
- [39] Stenger B. Model- based hand tracking using a hierarchical Bayesian filter. Phd thesis. University of Cambridge. 2004.
- [40] Hiramatsu T, Ogawa T, Haseyama M. A Kalman filter based restoration method for in- vehicle camera images in foggy conditions [C]. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, 1245-1248.
- [41] Joshi K R, Kamath R S. Quantification of retinex in enhancement of weather degraded images[C]. In: *Proceedings of IEEE International Conference on Audio, Language and Image Processing*, 2008, 1229-1233.
- [42] Kopf J, Neubert B, Chen B, et al. Deep photo: Model-based photograph enhancement and viewing. *ACM Transactions on Graphics*, 2008, 27(5) : 1-10.
- [43] Tan R T. Visibility in bad weather from a single image [C]. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, 2008: 1-8.
- [44] Rao Y, Chen L. A survey of video enhancement techniques [J]. *Journal of Information Hiding and Multimedia Signal Processing*, 2012, 3(1) : 71-99.
- [45] Ramesh R, Adrian I, Yu J. Image fusion for context enhancement and video surrealism. *ACM SIGGRAPH 2005 Courses*, 2005: 1-8.
- [46] Li J. Combining scene model and fusion for night video enhancement[J]. *Journal of Electronics*, 2009, 26 (1) : 88-93.
- [47] Loza A. Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients. *Digital Signal Processing*, 2013, 23(6) : 1856-1866.
- [48] Garg K, Nayar S K. Vision and Rain[J]. *Internet Journal of Computer Vision*, 2007, 75(1) : 3-27.
- [49] Barnum P, Narasimhan S, Barnum T K. Analysis of rain and snow in frequency Space[J]. *International Journal of Computer Vision*. 2010, 86(2) : 256-274.
- [50] Bossu J, Hautiere N, TarelJ P. Rain or snow detection in image sequences through use of a histogram orientation streaks[J]. *International Journal of Computer Vision*. 2011, 93(3) : 348-367.
- [51] Schmid U. , Schelten K, Roth S. Bayesian Deblurring with Integrated Noise Estimation. In: *Proceedings of CVPR*, 2011: 2625-2632.
- [52] Cho T S, Paris S, Berthold K P Horn. Blur Kernel Estimation using the Radon Transform. In: *Proceedings of CVPR*, 2011: 241-248.

- [53] Kim T H, Lee K M. Segmentation-Free Dynamic Scene Deblurring. In: Proceedings of CVPR, 2014, 2766-2773.
- [54] Glasner D, Bagon S, Irani M. Super-resolution from a single image. In: Proceedings of ICCV, 2009, 977-984.
- [55] Shahar O, Faktor A, Irani M. Space-time super-resolution from a single video. In: Proceedings of CVPR, 2011, 3353-3360.
- [56] Poppe R. A survey on vision-based human action recognition. *Image Visual. Computing*, 2010, 28(6): 976-990.
- [57] Weinland D, Ronfard R, Boyer E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Underst.*, 2011, 115(2): 224-241.
- [58] Popoola O P, Wang K. Video-Based Abnormal Human Behavior Recognition - A Review. *IEEE Trans. Syst. Man, Cybern. Part C Appl. Rev.*, 2012, 42(6): 865-878.
- [59] Ke S R, Thuc H, Lee Y J, et al. A Review on Video-Based Human Activity Recognition. *Computers*, 2013, 2(2): 88-131.
- [60] Xia L, Chen C C, Aggarwal J K. Human detection using depth information by Kinect [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2011, 15-22.
- [61] Smisek J, Jancosek M, Pajdla T. 3D with Kinect [M]. Consumer Depth Cameras for Computer Vision, Springer London, 2013: 3-25.
- [62] Lan T, Wang Y, Yang W, et al. Discriminative latent models for recognizing contextual group activities. *IEEE TPAMI*, 2012, 34(8): 1549-1562.
- [63] Fathi A, Rehg J M. Modeling Actions through State Changes [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2579-2586.
- [64] Prest A, Ferrari V, Schmid C. Explicit modeling of human-object interactions in realistic videos. *IEEE TPAMI*, 2013, 35(4): 835-48.
- [65] Wang X, Ji Q. Incorporating contextual knowledge to Dynamic Bayesian Networks for event recognition [C]. In: Proceedings of International Conference on Pattern Recognition, 2012: 3378-3381.
- [66] Song Y, Morency L P, Davis R. Action Recognition by Hierarchical Sequence Summarization [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3562-3569.
- [67] Boiman O, Irani M. Detecting Irregularities in Images and in Video [J]. *International Journal of Computer Vision*, 2007, 74(1): 17-31.
- [68] Cong Y, Yuan J, Liu J. Sparse reconstruction cost for abnormal event detection [J]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011: 3449-3456.
- [69] Zhao B, Li F F, Xing E P. Online detection of unusual events in videos via dynamic sparse coding [J]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2011: 3313-3320.
- [70] Calderara S, Cucchiara R, Prati A. Detection of abnormal behaviors using a mixture of Von Mises distributions [C]. In: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, 2007: 141-146.
- [71] Ballan L, Bertini M, Bimbo A D, et al. Effective Codebooks for human action categorization [C]. In: Proceedings of IEEE International Conference on Computer Vision Workshops, 2009: 506-513.
- [72] Kejun W, Popoola O P. Ant-based clustering of visual-words for unsupervised human action recognition. In: Proceedings of World Congress on Nature and Biologically Inspired Computing, 2010: 654-659.

- [73] Kratz L, Nishino K. Anomaly detection in extremely crowded scenes using spatiotemporal motion pattern models [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, 1446-1453.
- [74] Kim J, Grauman K. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009: 2921-2928.
- [75] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model [C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009, 935-942.
- [76] Varadarajan J, Emonet R, Odobej J M. A Sequential Topic Model for Mining Recurrent Activities from Long Term Video Logs[J]. International Journal of Computer Vision, 2012, 103(1): 100-126.
- [77] ISO/IEC JTC1/SC29/WG11 (MPEG). N6231 “Report of The Formal Verification Tests on AVC (ISO/IEC 14496-10 + ITU-T Rec. H.264)”. 2003, Waikoloa.
- [78] 国家广播电影电视总局广播电视台规划院. AVS 视频压缩质量主观评价(AVS 参考软件 5.2 版)测试报告. 2005.
- [79] Zhang X, Liang L, Huang Q, et al. An Efficient Coding Scheme for Surveillance Videos Captured by Stationary Cameras. In: Proceedings of Visual Communication and Image Processing, 2010.
- [80] Zhang X, Huang T, Tian Y, et al. Low- Complexity and High- Efficiency Background Modeling for Surveillance Video Coding. In: Proceedings of Visual Communication and Image Processing, 2010: 1-8.
- [81] X Zhang, T Huang, Y Tian, Hierarchical- and- Adaptive Bit- allocation with Selective Background Prediction for High Efficiency Video Coding(HEVC), In: Proceedings of Data Compression, 2013: 1-8.
- [82] Zhang X, Tian Y, Huang T, et al. Macro- block- level Selective Background Difference Coding for Surveillance Video[C]. In: Proceedings of IEEE International Conference on Multimedia Expo, 2012: 1067-1072.
- [83] Li Y, Ai H, Lao S. Tracking in Low Frame Rate Video: A Cascade Particle Filter with Discriminative Observers of Different Lifespans[C]. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2007: 1-8.
- [84] Zheng Y, Jiang Y G, Xue X. Learning hybrid part filters for scene recognition. In: Proceeding s of Computer Vision-ECCV, 2012: 172-185.
- [85] Jiang Y G, Dai Q, Xue X, et al. Trajectory- based modeling of human actions with motion reference points. In: Proceedings of Computer Vision-ECCV. 2012: 425-438.
- [86] He K, Sun J, Tang X. Single image haze removal using dark channel prior. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2341-2353.
- [87] Meng G, Wang Y, Duan J, et al. Efficient Image Dehazing with Boundary Constraint and Contextual Regularization[C]. In: Proceedings of IEEE International Conference on Computer Vision, 2013: 617-624.
- [88] Rao Yunbo. An effecive night video enhancement algorithm[C]. In: Proceedings of IEEE Conference on Visual Communications and Image Processing(VCIP) , 2011: 1-4.
- [89] Rao Y, Lin W, Chen L. Image- based fusion for video enhancement of night- time surveillance. Optical Engineering, 2010: 1-10.
- [90] Fu H, Ma H, Wu S. Night Removal by Color Estimation and sparse representation[C]. In: Proceedings of International Conference on Pattern Recognition(ICPR) , 2012: 3356-2259.
- [91] 胡巍, 何小海, 高明亮, 等. 一种新型的雨雪视频图像复原方法[J]. 四川大学学报, 2012, 44

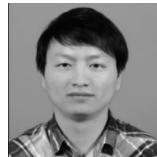
- (1): 1-5.
- [92] 孙毅刚, 段晓晔, 张红颖, 于之靖. 基于改进 snake 模型的图像中雨雪去除算法研究[J]. 计算机应用研究, 2011, 28(5): 1-3.
- [93] 张小红. 视频去运动模糊及超分辨率研究[D]. 杭州: 浙江大学图书馆, 2012.
- [94] 唐磊. 多帧图像超分辨率重建算法研究[D]. 杭州: 浙江大学图书馆, 2011.
- [95] Wang H, Kläser A, Schmid C, et al. Dense Trajectories and Motion Boundary Descriptors for Action Recognition[J]. International Journal of Computer Vision, 2013, 103(1): 60-79.
- [96] Jiang Y G, Dai Q, Xue X, et al. Trajectory-Based Modeling of Human Actions with Motion Reference Points. In: Proceedings of Computer Vision-ECCV, 2012: 425-438.
- [97] Kong Y, Jia Y. A Hierarchical Model for Human Interaction Recognition[C]. In: Proceedings of IEEE International Conference on Multimedia and Expo, 2012: 1-6.
- [98] Ouyang W, Wang X. Single-Pedestrian Detection Aided by Multi-pedestrian Detection[C]. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2013: 3198-3205.
- [99] Sun C, Zhang T, Bao B K, et al. Latent support vector machine for sign language recognition with Kinect [C]. In: Proceedings of IEEE International Conference on Image Processing, 2013: 4190-4194.
- [100] Lu C, Shi J, Jia J. Abnormal Event Detection at 150 FPS in MATLAB[C]. In: Proceedings of IEEE International Conference on Computer Vision, 2013: 2720-2727.
- [101] Jiang Y G, Bhattacharya S, Chang S F, et al. High-level event recognition in unconstrained videos[J]. International Journal of Multimedia Information Retrieval, 2012, 2(2): 73-101.
- [102] 谌先敢, 刘娟, 高智勇, 等. 基于累积边缘图像的现实人体动作识别[J]. 自动化学报, 2012, 38(8): 1380-1384.
- [103] 郭梓鑫, 衣杨, 李汉臣. 基于自适应特征融合的自然环境视频行为识别[J]. 计算机学报, 2013, 36(11): 2330-2339.
- [104] 崔永艳, 高阳. 基于多示例学习的异常行为检测方法[J]. 模式识别与人工智能, 2012, 24(6): 862-868.
- [105] 朱旭东, 刘志镜. 基于主题隐马尔可夫模型的人体异常行为识别[J]. 计算机科学, 2012, 39(3): 251-255.
- [106] 覃勋辉, 王修飞, 周曦, 等. 多种人群密度场景下的人群计数[J]. 中国图象图形学报, 2013, 18(4): 392-789.
- [107] Goyette N, Jodoin P M, Porikli F, et al. Chagedetection.net: A new change detection benchmark dataset[C]. In: Proceedings of, 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops(CVPRW), 2012: 1-8.
- [108] Wu Y, Lim, Yang M H. Online object tracking: A benchmark. In: Proceedings of CVPR, 2013, 2411-2418.

作者简介

本报告的“监控视频编码”由北京大学黄铁军撰写,“运动目标检测与跟踪”由北京航空航天大学的李波和郑锦撰写,“监控视频增强”由北京邮电大学的马华东和傅慧

源撰写，“视频动作与异常行为识别”由复旦大学的薛向阳和姜育刚撰写，报告由华中科技大学的于俊清负责策划和统稿。以下以姓氏拼音排序。

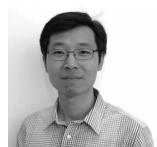
傅慧源 博士，北京邮电大学计算机学院讲师。主要研究方向为视频大数据、计算机视觉与模式识别、多媒体技术等。CCF 会员。



黄铁军 博士，北京大学信息科学技术学院数字媒体研究所所长，教授，主要研究方向为图像识别和视频编码，tjhuang@pku.edu.cn。



姜育刚 博士，复旦大学计算机科学技术学院副教授，博士生导师，CCF 会员。主要研究方向为多媒体内容分析和计算机视觉。



李波 博士，北京航空航天大学计算机学院教授，博士生导师。主要研究方向包括图像压缩与传输、视频分析与理解、遥感信息融合和嵌入式多媒体系统。中国计算机学会多媒体专业委员会副主任。



马华东 博士，北京邮电大学计算机学院教授、博士生导师。主要研究方向为多媒体系统与网络、物联网等。中国计算机学会多媒体专业委员会副主任。



薛向阳，博士，复旦大学计算机科学技术学院教授、博士生导师，CCF 杰出会员。主要研究方向为视频图像内容分析与检索、计算机视觉等。



于俊清 博士，华中科技大学计算机科学与技术学院教授，博士生导师，主要研究方向为数字媒体处理与检索、多核计算与流编译。中国计算机学会多媒体专业委员会副主任。



郑 锦 博士，北京航空航天大学计算机学院讲师，硕士生导师，主要研究方向为视频图像分析处理、运动目标检测跟踪。



穿戴式计算研究进展与趋势

CCF 普适计算专业委员会

李石坚¹ 班晓娟² 叶振宇¹ 沈 晴² 潘 纲¹

¹浙江大学计算机学院，杭州

²北京科技大学计算机与通信工程学院，北京

摘要

与 PC 和智能手机相比，穿戴式设备具有更好的随身服务特性、更强的感知能力、更便利的交互功能，正成为信息服务的重要载体。伴随传感、材料、交互技术的进步，穿戴式设备的形态、功能、服务和应用模式迅速发展。本文从传感技术、计算模型、交互技术等方面综述了穿戴式计算的国内外进展，对国内外进展进行了对比，并分析了穿戴式计算的发展趋势和挑战。

关键词：穿戴式计算，普适计算，人机交互

Abstract

Comparing with PC and Smartphone, the wearable devices have many advantages, such as better portability, more powerful sensory ability and more convenient interaction. Today, the wearable devices become a kind of important platform for information services. Meanwhile, with the technical progress of sensors, material and human-computer interaction, the wearable devices have made rapidly growth. This report analyzes the state-of-art about wearable computing, and comparing the development status at home and abroad. Finally, the report discusses the future trends and the potential challenges of wearable computing.

KeyWords: Wearable Computing, Ubiquitous Computing, Human-computer interaction

1 引言

20世纪90年代，Mark Weiser提出了“普适计算”的概念，认为计算机将退居幕后直至消失，随之而来的是人们获得“无处不在”的信息服务。

近30年来，个人计算机（Personal Computer, PC）作为个人信息服务的主要载体和入口，构成了个人信息环境的核心部分。然而今时今日，随着智能手机和各类穿戴设备等智能终端的兴起，大量原本由PC承担的信息服务功能转移到这些终端上来。并且，由于这些终端具有更好的随身服务特性、更强的感知能力和更便利的交互功能，衍生出了许多PC无法承担的信息服务。随之而来的是PC计算机的销量持续下滑，据国际数据公

司 IDC 数据显示，2013 第一季度全球 PC 销量下滑 14%，为近二十年来最大降幅。可见，手机和各类穿戴设备等异构终端正在成为信息服务的主要载体，计算机的“消失”正逐渐成为现实。

在穿戴设备方兴未艾之时，新传感、新材料、新的基础设施（云和物联网）正在推动穿戴设备的形态、作用发生巨大变化。穿戴计算的转型和升级正在加速进行。我们认为新的技术进步对穿戴计算的影响会体现为：1) 依托云平台和各类互联技术，设备不断隐藏到环境中，特别是柔性电子等技术使得无间断的计算、感知和显示技术成为可能，穿戴设备将逐渐“融入”到基础设施中；2) 源于物联网、社交网等渠道的海量数据，使得基于各类预判结果的推送式计算服务不断出现，用户对无时无刻佩戴各类设备的需求被降低。3) 各种无穿戴、非接触的新型交互手段迅速发展，体感交互、语音交互等技术让用户脱离穿戴设备依然能得到计算服务。

围绕上述趋势，本报告将对国内外近年来在穿戴计算领域的重要进展进行小结和对比分析，并试展望该领域未来趋势及相应挑战。

2 国际研究现状

2.1 穿戴式计算国际研究现状

穿戴式计算诞生于 20 世纪 70 年代，最早源自加拿大多伦多大学的 Steve Mann 教授^[1]，其发明的用于控制照相设备的穿戴式计算机是第一个真正意义上的穿戴式产品（图 1）^[2]。在此之后，穿戴式计算随着各项相关技术（包括语音识别、传感器、无线通讯、电池等）的不断进步而迅猛发展，涌现出许多新型产品和服务，市场快速增长，成为多学科、多领域、多行业关注的热点。



图 1 穿戴式计算机先驱者 Steve Mann 的研究历程

2.1.1 关键技术研究现状

穿戴设备形态和功能各异，但其核心能力大致可分为数据传感和数据分析两部分。同时，在这两方面关键技术的支持下，衍生出了一系列新的应用模式。

2.1.1.1 传感技术

传感数据是穿戴式设备的主要输入，是后续分析、计算和服务的依据。穿戴式设备通过各种不同类型的传感器获取来自环境、人体等不同来源的信息，进行加工、存储和分析，进而支持上层应用服务和交互行为。

(1) MEMS 传感器

微机电系统（microelectro mechanical system, MEMS）传感器是采用微机械加工技术制造的新型传感器（图 2），包括微加速度传感器、微机械陀螺仪以及微磁传感器等，主要用于医疗、汽车电子和运动追踪系统等方面^[3]。目前已有多款较为适合穿戴计算需求的传感器推出。例如，Bosch 针对穿戴式设备的特殊需求提出了多种不同的解决方案（图 3），包括独立式的三轴加速度传感器、三轴陀螺仪，也包括了同时具备加速度计、陀螺仪、磁场的 9 轴传感器。后一类传感器的体积达到 2mm 的级别，功耗降低到 uA 的级别，而且可以将多种类型的传感器、传感器微控制芯片以及传感数据预处理程序都整合在一个模块上。例如 Bosch 的 BME280 传感器，集合了压力、温度、湿度传感功能，长宽为 2.5mm，功耗电流 15uA，在尺寸、能耗、精度等方面较适合穿戴式计算需求。

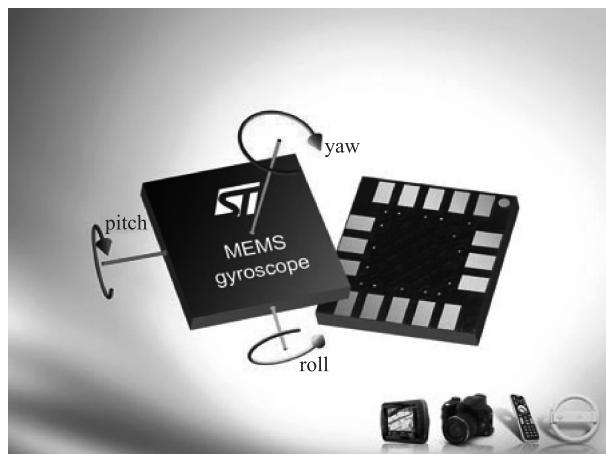


图 2 微机电系统传感器及其典型应用

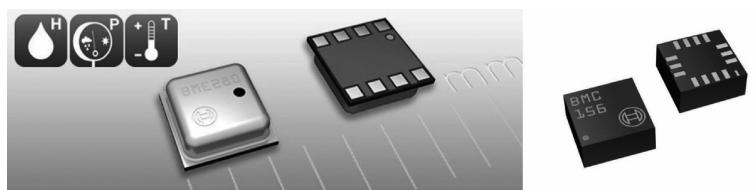


图 3 Bosch 的各种系列传感器芯片

这些日益成熟的 MEMS 传感器已经被运用到了不少市场化或准市场化的穿戴式设备上（图 4），如 Nike FuelBand 腕带就内置了微加速度传感器等，用于记录用户的运动数据。一些头戴式显示器也用到了 9 轴 MEMS 传感器，用于方向识别和位置确定。现阶段各类 MEMS 传感器是穿戴式设备上应用最广泛的核心技术。



图 4 使用了 9 轴 MEMS 传感器的头戴式显示器与运动手环

（2）生理信息传感

穿戴式设备被紧密部署在人体上，非常适合用于实时采集人体的生理状况信息，在医疗、运动等领域有广泛的前景。目前，应用较多的穿戴式生理信息传感设备有两类^[4]，一类为传统的个人随身物品，如腕表、臂环、手机、指环等；另一类为电子织物，如各种基于无线传感网络的智能衫等（图 5）^[5]。



图 5 智能传感 T 恤 D-Shirt 及与之互联的手机

目前国际上已经有利用穿戴式设备检测血糖、血压、血氧等重要生理信息比较成熟的方案。例如，美国 Cygnus 公司生产的手表式血糖仪，利用电场来刺激皮肤，通过皮肤的渗透作用，使一部分血糖分子穿过皮肤集聚在电极周围，然后用生化方法测量血糖参数。美国 Medwave 公司开发出一款名为 Vasotrac 的腕式血压测量仪，通过周期性地在桡

动脉上加压和减压来确定血管零负荷（Zero load）状态，并在该状态下通过脉搏波的幅值和从波形中提取的其他参数来确定血压值。美国 SPO Medical 公司推出一款“血氧手表”，可以在睡眠过程中监测使用者的血氧饱和度，从而降低患有睡眠窒息症者在夜间出现呼吸阻碍的危险。当然这些生理信息的检测方法仍需要进一步的研究，部分方法仍然存在需要微创、外界干预过多等弊端。

利用电子织物适配传感器检测生理信号也得到了越来越广泛的运用。这类技术有两种实现方式，一种是将传统的传感器，如微控制器、发光二极管、光纤和压电传感器等集成到布料中；另一种是开发基于有机材料，即电活性聚合物（EAP）的装置。新的纺织技术，如将金属丝编织到织物中去的技术的发展，加速了采用第一种方法制成的电子织物在电连接、数据通讯和供电等方面的应用。美国佐治亚理工学院、美国 VivoMetrics 公司和法国的一个研究小组在其进行的研究中均采用了此种技术。

（3）深度传感技术

近年来 3D 环境信息的采集技术得到了长足发展。目前常用的摄像头获取的都是平面 2D 信息，要获取 3D 信息则需要额外的景深摄像头，即通过两个摄像头同时获取场景图片，再经计算获得类似人眼采集信息的效果。

深度传感技术为穿戴设备提供了更丰富的输入信息，例如 OmniTouch 穿戴式多点触控投影系统（图 6 左）就应用了深度传感技术。OmniTouch 通过深度传感器来捕捉用户的操作，配合激光投影仪将图形界面投射于物体表面，从而将任何表面变为“可触控”界面，以便支持多点触控和多种手势等操作。Meta 的智能眼镜 SpaceGlasses（图 6 右）同样是一款已经使用了深度传感技术的可穿戴设备，其利用景深摄像头实现对深度的识别，据此可进一步识别手势动作，从而进行增强现实的游戏、工艺制作等。



图 6 使用了深度传感技术的穿戴式设备

（4）生物电^[6]和骨传导

生物体在生命活动过程中，其器官、组织和细胞会发生电位和极性的变化，依据这种变化有可能分析出相应的生理活动。如 MYO 腕带（图 7 左）能捕捉到用户手臂肌肉运动时产生的生物电变化，从而识别佩戴者的手势，进而实现利用不同手勢动作来操控设备的功能。

骨传导技术是一种利用声音能通过头骨、颌骨传到听觉神经并引起听觉的传导技术。通过骨传导，人们可以感知到频率高达 120kHz 的超声波。Orb 蓝牙耳机（图 7 右）就应用了这一技术，当接听来电时，将耳机挂在耳朵上就能实现通话，而不用像普通耳机那样塞进耳朵里。使用骨传导技术可以使双耳获得自由，不再受到束缚。



图 7 MYO 腕带与 Orb 蓝牙耳机

2.1.1.2 数据分析处理和行为识别

通过各种传感手段采集到传感数据后，如何从数据中进行分析，以提取终端用户或者数据服务商感兴趣的内容也是穿戴式计算的重要研究方向，是提供各种应用服务的依据。

某些简单信息只需要做存储和可视化，即可满足用户需求，如温度趋势变化，温度预测等。如美国佐治亚理工学院的智慧衫（Smart Skirt）其在布料生产过程中，作为数据总线的金属纤维和柔性光纤呈螺旋状被织在布料里。传感器可以插入与光纤相连的 T 型连接器，通过与数据总线相连的 T 型连接器把检测到的生理信号传输给监测装置。这种棉质智慧衫可以监测心率、呼吸、体温及其他生理参数。该项发明已由美国 Sensatex 公司作进一步开发并推向市场（图 8）。

对于加速度计、陀螺仪产生的运动信息，则需要对信息进行模式识别、分类等处理，以获得如用户手势，用户运动状态等信息。这方面的应用主要集中在各种智能手表上，如 Nike + 跑鞋，结合 Nike + iPod 运动套件或传感器，通过对各种运动传感器采集到的实时运动数据进行分析，从而获得运动时间、距离、速度等信息，便于用户分析自己的运动状态。另一方面，这类运动传感器支持用户手势的操作，从而可以用更自然的交互方式控制可穿戴设备。三星在 US2014/0139422A1 专利^[7] 中披露了自己的智能手表内置若干传感器，支持一系列手势操作（图 9）。

深度传感数据的处理更为复杂。一般原始数据来自两个摄像头，如 META 的智能眼镜，其配有两个摄像头，一个为颜色摄像头，用于获得色彩等信息；一个为景深摄像头，用于获得距离信息。在采集到摄像头记录的手指、手臂的信息后，可分析出其空间位置。不仅能够识别传统的抓取、指向、捏取等，还可以识别出更为复杂的手势，如滑动、推动等。将这些复杂的手势识别结果和虚拟现实技术结合，可以完成增强现实游戏、模拟工艺设计（图 10）等诸多功能。

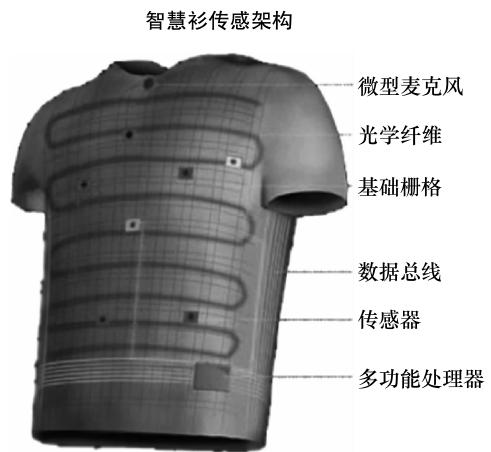


图 8 Sensatex 公司的智慧衫

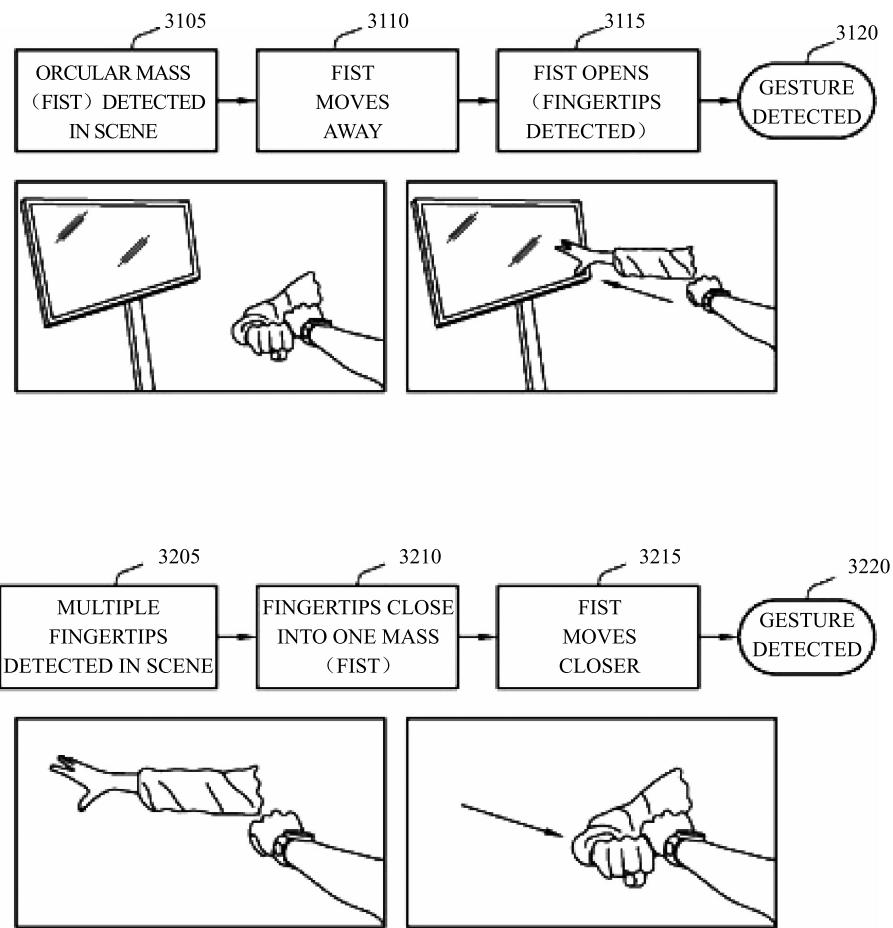


图 9 三星专利中的手势示意图



图 10 使用深度传感数据识别手势进行工艺模拟设计

2.1.1.3 穿戴设备计算模型

受限于设备本身的体积和舒适性等方面的要求，穿戴设备的硬件资源相对有限，因而造成其功能高度专一，难以拓展与延伸。如果单纯依赖设备自身，其功能和应用场景会大受限制。得益于无线通信技术的发展和云平台技术的发展，将穿戴式计算和云计算相结合，穿戴式设备的应用前景变得十分广阔。

现阶段穿戴式设备产品的成功很大程度上取决于其所依赖的平台和应用生态系统是否完善。以谷歌眼镜为例，谷歌工程师团队开发的眼镜一开始就和谷歌的服务捆绑在一起，依托谷歌庞大的开发者群体、完善的开发工具、应用软件商店渠道及数以百万计的应用软件，使得谷歌眼镜具有巨大的先发优势。2013年第二季度，谷歌宣布成立“眼镜联盟（Glass Collective）”，旨在进一步完善谷歌眼镜专有的生态体系，培养谷歌眼镜的开发者，并帮助谷歌眼镜以最快速度成为大众主流设备（图11）。

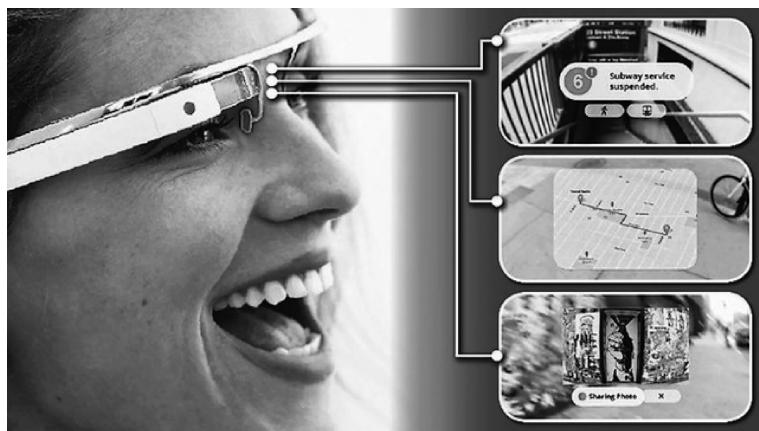


图11 谷歌眼镜结合云计算，提供各类服务

产业界其他巨头推出的穿戴设备也采取了类似思路。例如，Nike + Fuelband 腕表不仅可以让人一边听线上音乐，一边跑步，还能收集用户的各项身体数据，并将其上传到耐克的数字平台“Nike +”上进行人体健康分析，然后与用户在 Nike + 上的其他数据横向融合。Nike + Fuelband 是耐克向运动数字化产品转型的标志之一，将“跑步”与“音乐”结合在一起，并以“数据”相连，最终汇集到平台之上。此外，耐克公司还部署了耐克 iPod 运动套装、Nike + 芯片、Nike + GPS、Nike + FuelBand，形成了耐克独有的生态体系，汇聚了大量的用户个性化数据。

迪士尼公司也正利用穿戴式技术的发展打造自己的平台和改变自己的经营方式。迪士尼乐园推出的 MagicBands（图12）以无线射频识别（Radio Frequency Identification, RFID）技术标记用户，并将每位顾客资料传送至云端。数据内容包括顾客目前所在的位置、购买纪念品数量等。通过分析穿戴设备获取的数据，迪士尼可以针对不同的客户群改善服务，从而提高迪士尼乐园的效益，如可分析在乐园里遇到“灰姑娘”的顾客，到了纪念品商店会不会购买她的衣服等。



图 12 在第 11 届 All Things Digital 大会上，迪士尼推出自己的 MagicBand

从上述领军企业的思路可以预见，穿戴式计算最终会从单一设备衍生出类似智能手机生态系统的格局。当前主流的穿戴设备计算模式大致如下：穿戴式设备可以充分利用自身强大的感知能力，获取各种类型的数据，这些数据可以是环境相关也可以是人体相关，对这些数据进行预处理后，可以在穿戴式设备自身进行相关的功能呈现，也可以将数据传输到云端，结合强大的云计算和云存储，从海量传感数据中挖掘出有价值的信息，既可据此对用户进行反馈和服务，亦可为其他决策做出依据。

另一方面，随着穿戴式设备的不断普及，各类穿戴式设备间以及与其他智能设备间的协同也是穿戴式计算模型中重要的一部分。不同的穿戴式设备间的协同可以弥补单一设备功能高度专一的弊端，改善系统整体的感知和服务能力。比如耐克的智能鞋和智能胸带可以协同工作，智能鞋可以用于采集运动信息，如运动速度、距离、时间等，而智能胸带可以采集心率信息，两者相结合可以让运动者了解自己的运动强度，以免造成运动损伤。此外，穿戴式设备可以和智能手机结合，从而大为拓展穿戴式系统的计算、存储和分析能力。一方面，智能手机可以存储穿戴式设备采集的各种信息，用于小规模的数据分析；另一方面，通过手机屏幕能够在移动环境中更好地可视化各种传感信息。至于穿戴式设备和其他设备的协同案例也已出现在市场上，如咕咚智能运动手环和智能体重计间协同，可以监测使用者的运动减肥效果等。

2.1.2 典型产品和服务

目前在国际上已经推出了不少穿戴式设备产品和服务，涵盖了运动、医疗、健康、出行、游戏等各种领域。这些产品中，有很少几类是完全独立工作，不能与其他设备协同。绝大部分产品都依托云计算平台，并与其他设备协同工作完成更复杂的任务。表 1 总结了比较典型的一些设备形态及其代表性产品。

表 1 一些典型穿戴式设备形态及其代表性产品 [6]

设备形态	设备名称
腕带、手环	MYO 腕带、Amiigo 腕带、Jawbone Up 2、Relay 等
手表	GEAK Watch、Pebble、Sony SmartWatch、iWatch 等
戒指	GEAK Ring 等
手套	Glove One 等

(续)

设备形态	设备名称
头箍	NeuroSky MindWave、Emotive EEG 等
鞋子	谷歌“会说话的鞋子”(Google talking shoe)、“何处是家园”(No Place Like Home) 卫星导航鞋等
耳环	Orb 蓝牙耳环等
挂件	Memoto 微型自动相机等
眼镜	谷歌眼镜、EyePhone 等
衣服	T 恤鼓点机 (Make beats)、美国麻省理工学院 Media Lab 音乐夹克衫等
裤子	嵌上键盘的裤子等
全身	SixSense、Wireless Body Area Network、Xpod 等

2.2 后穿戴计算：非接触式交互技术

在普适计算的时代，一个人将同时享受成百上千计算机的服务，计算机技术将朝着以人为本，解放用户、增强体感的方向继续发展^[8]。这个时期要求对服务进行过滤，以期适应人的生活和工作。围绕这一愿景的一个重要趋势是：交互设备将逐渐隐藏于环境之中，交互的深度和广度不断扩展，不用人们去关注服务的实现细节，甚至在没有觉察到设备或技术存在时即可享受到各类服务。

围绕这一目标，当前的人机交互范式力图在日常生活和工作中摆脱设备的束缚，借助各种传感器、图像、语音采集设备，通过手势、语言、体态等方式随时随地与计算机进行交互^[9]。环境中的大量传感器将不断感知环境变化和用户变化，自动根据相关的上下文来为用户提供服务^[10]。这种思想对穿戴计算的交互模式产生了深刻影响，穿戴式设备的交互方式也日益朝着普适、自然、非接触或弱接触的方向发展，力图不断将人从设备中“解放”出来。

从近年来的研究发展上看，部署在环境中的非接触或弱接触式传感装置在种类上较为发散，语音、手势、眼动、脑电等感知手段衍生出大量的交互方法，部分已广泛应用。同时，从普通的摄像机镜头和麦克风发展出来的多种类别（光学、电磁、超声等）的传感器，可全方位地捕捉参与者的姿态、动作、表情、声音，再进一步结合情感计算、虚拟现实和增强现实等技术，打造出具有更强沉浸感的交互方式。可以预见，这些新型交互手段有可能部分取代现有穿戴设备的交互功能，使得各类穿戴计算设备不断小型、微型化直至完全消失。

2.2.1 交互方式的新进展

(1) 体感交互

基于光学传感器进行深度获取，进而实现体感人机互动，是近年来人机交互等领域的研究热点。

微软的 Kinect 体感装置是这方面的代表性成果，其特点是结合了摄像头和红外深度探测器，能够以相对较好的实时性获取深度图像序列，实现 3D 数据捕捉、人体识别和骨骼跟踪，

并可进一步实现各种体感互动应用。初始作为体感游戏控制器而出现的 Kinect，不仅仅改变了游戏，同时也改变了交互设计中计算机和使用者的角色。计算机对用户感知能力被进一步增强，使得传统的图形交互界面的统治被撼动，计算机直接与用户进行交流成为可能。

Leap Motion 则是通过高精度的红外传感器感知其上方的手指动作，从而可精确判断出铅笔和手指的区别。Leap Motion 在手势识别的应用领域有着十分广阔前景。



图 13 Kinect 与 Leap Motion

除了基于视频或红外设备的手势和姿态识别技术，利用加速度传感器等微型传感设备的行为识别也被广泛用于移动终端和智能电视。浙江大学提出了一种称为 MagicPhone^[11]的智能手机原型，利用各种传感器识别用户的指向、滑动等手势行为，并用于与周围环境交互，例如用指向和手势操控家电等。三星已推出具有眼球追踪和手势识别功能的新一代手机，并尝试将这些技术应用于其最新的液晶电视。

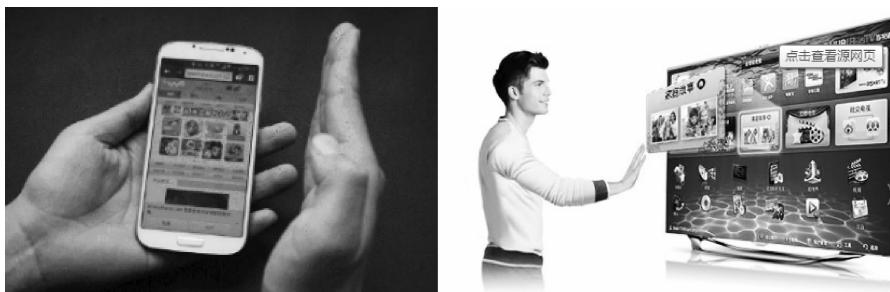


图 14 手势控制手机与电视

除了上述体感输入方法外，基于超声波及多普勒效应等的运动检测方法也已被研究人员广泛用于非接触式体感输入的研究中。此外，由于各类电磁信号广泛充斥于我们的生活，有研究者通过分析电场信号、电器以及电子设备产生的电磁噪声信号被用户行为扰动的模式，进而识别用户的行为，这也是一类重要的非接触体感交互方式。

(2) 语音交互

由于语言可以承载人类大部分意图的表达，若通过说话就可以得到计算机的反馈和服务，将使人与计算机的交互更为直接和自然。语音交互在穿戴式计算领域乃至整个人机交互领域有着极为重要的地位^[12]。

苹果公司的 Siri 是语音交互技术非常成功的一项应用。在结合了自然语言理解、知识库等技术的基础上，Siri 将移动终端设备成为了一台智能机器人。Siri 可以理解用户对一些

服务的请求，如：阅读短信，介绍餐厅、询问天气、设置闹钟等等。Siri 最大的特色是不仅有十分生动的对话接口，还能够不断学习新的声音和语调，提供对话式的应答。例如使用者说出的内容如果包括了“喝了点”、“家”这些字（甚至不需要符合语法，相当人性化），Siri 则会判断当前用户意图是醉酒后要回家，将自动给出建议是否要帮忙叫出租车。



图 15 苹果语音助手 Siri

(3) 脑机交互

如何理解人的意图是神经科学、计算机、心理等学科的重要交叉研究课题，该领域近年的进展带来了一系列新型交互模式。脑机交互是这类人机交互模式的代表，显示出巨大的潜力。

InteraXon 公司推出了一款脑电波感应头带，佩戴上这款头带，用户可以通过脑电波来控制窗帘以及灯光。InteraXon 的脑电波感应头带采用脑机接口技术，能够解读用户的专注程度。基于此项技术推出的意念游戏机也已进入大众的视野。

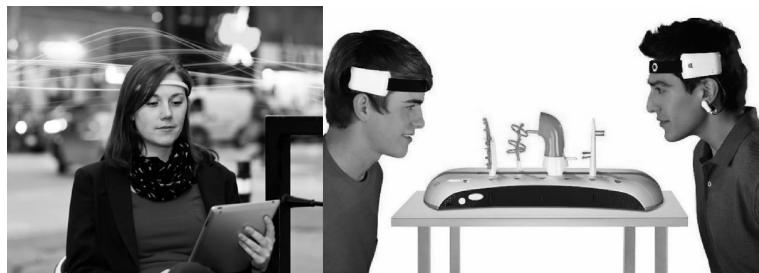


图 16 脑电波感应头带

(4) 触感交互——柔性电子技术

柔性电子（亦称为“软电子”）技术是近些年的研究热点，在 2014 年的“十大世界创新科技技术”就有 6 个和柔性电子有关。目前市场上已经出现了大量基于柔性电子的穿戴式产品，涵盖腕表、腕带、眼镜、智能织物、电子皮肤等。其基础元件包括柔性屏幕、柔性电池、柔性传感阵列等。

从交互角度来看，基于柔性传感阵列的电子皮肤系统可提供类似人体皮肤般的触感交互能力。传统的传感设备都是对单一点进行传感数据的采集，或者简单地由多个传感器构成传感阵列。但这种构建方式过于粗糙，很难部署到形态多变、大小不一的复杂环境中，体积也很难微型化，不适合穿戴计算需求。为改善这一状况，研究者开始尝试使用微小的

柔性传感阵列贴片，通过高密度、可变形的传感阵列感知压力、温度等信息，并分析阵列采集的传感数据，提取姿态、触感等信息，据此进行交互和反馈^[13]。另一个研究的方向是电子纹身（表皮电子系统）^[14]，可通过印刷方式将电子电路像刺青一样粘在皮肤上，可以随意地拉伸、弯曲和旋转，通过电子电路上的传感器记录温度、压力等各类信息，甚至还集成了太阳能电池和通讯天线。目前，国际上较成熟的贴片触感解决方案有 Biotac^[15] 和 PPS，均已推出相关的触感阵列产品，较多应用于机器人触感增强方面。

这种基于触感的交互方式可带来全新的体验。目前已经商业化的产品如美国的 UGOBE 公司研发的 Pleo 电子宠物，通过覆盖在宠物表面的触觉传感器，使得用户可以通过抚摸等方式和宠物进行交互。

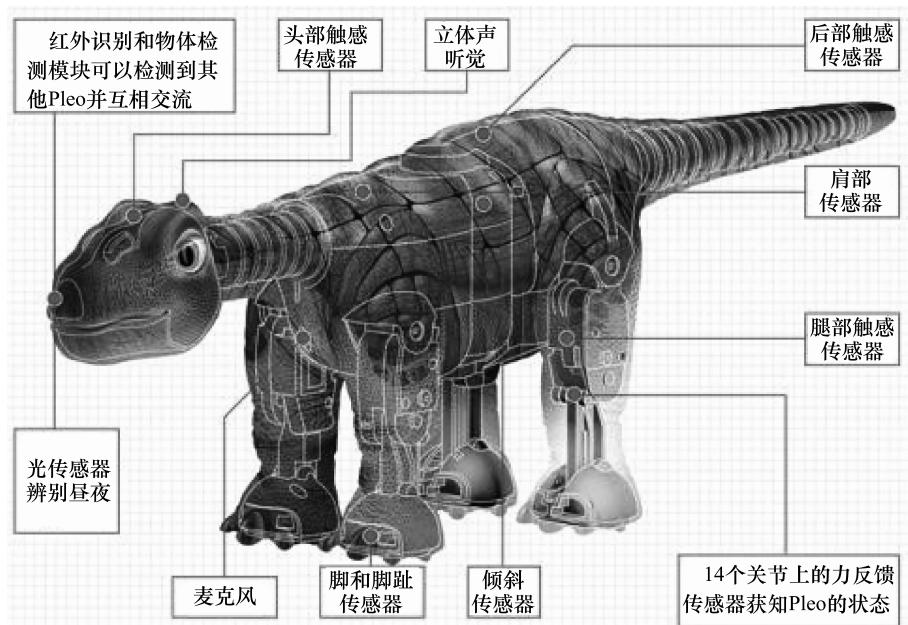


图 17 Pleo 电子宠物及各种传感器分布

2.2.2 交互设备逐渐隐藏

正如美国微软公司在《人类的本质：2020 年的人机交互》报告所称，“计算机和人类之间的生理界限将变得模糊……目前存在于人机界面之间牢固的界限正在消失，这种界面会离我们越来越近，电子装置将融入我们的衣服甚至身体，另一方面，计算机将悄悄融入我们的生活环境，例如，进入日用品之中。”

随着环境设备的不断增强，人机交互界面呈现出大屏化和可携带化两组方向的发展。随处可见的玻璃大屏和随身携带的智能设备，使得交互界面可以处于人们生活的任意角落。通过与这些设备的融合，非接触式的输入设备要么隐藏于环境，要么融合入“穿戴式”设备，朝着逐渐“消失”的趋势发展^[16]。同时，通过与情感计算、虚拟现实和增强现实技术的结合，新型的交互模式将带来更人性化的体验和更强的沉浸感。

情感能识别技术能够对理解人类情感的产生、表达和感知具有重要的价值，在人机交互方面具有广泛的应用前景^[17]。通过引入语音情感技术，机器人或口语对话系统能够更加自然地与人进行对话；通过引入表情和动作的情感分析，交互系统可更准确的判断人的意图，提供情感上的帮助和监控，从而变得更为人性化。

虚拟现实和增强现实技术则进一步隐藏了交互设备的存在感。增强现实技术将计算机生成的虚拟信息精准地叠加在真实环境中，虚拟现实技术则将用户融入虚拟世界里。新型的交互输入设备与各类大屏幕、投影仪、移动终端的结合，将使用户逐渐模糊了虚拟与现实的边界，完全融入到自然无障碍的交互中。

未来人机交互界面的发展趋势，体现了对人的因素的重视，标志着人机交互技术从“人适应计算机”向“计算机不断地适应人”方向发展。传统的界面事实上成为隔离物质世界和信息世界之间的屏障，虚拟现实、情感计算等技术的应用，将实现“虚物实化”和“实物虚化”，消除物理对象和抽象对象、输入装置和输出装置在交互空间中的差别，并为人提供多感觉通道的自然临境体验；语音及文字识别和自然语言理解等言语计算，手写体和手绘草图识别等笔式计算及手势和表情识别、视觉-目标拾取认知技术等视觉计算等技术的不断发展和完善将不断提高人机交互的智能化程度，使机器能够根据上下文及使用者的特点主动识别人的身姿、手势、语音和表情等各种自然行为，进而判断出人的意图。

MIT 的第六感项目^[18]所提出的一套交互模式和系统是这一趋势的典型代表。通过一系列输入输出设备的集成，该系统融合了多种交互方式，体现了现实空间和虚拟空间的“融合”。整个系统由四个套在手指上的彩色标记环、一个小型摄像头、一个便携式投影仪等设备组成。通过摄像头追踪标记环的运动，可以识别用户手势信息，进而结合当前的用户情境，推断出用户动作的意图，例如拍照、文本输入等。执行结果和操作界面可以通过投影仪投影到任何合适的显示位置，达到增强现实的效果。



图 18 第六感系统

谷歌眼镜（Google Glass）则是在这一发展趋势中当前最为成功的商业化产品。谷歌眼镜利用的是光学反射投影原理（HUD），首先将光投到一块反射屏上，而后通过一块凸透镜折射到人体眼球，在人眼前形成一个足够大的可以显示简单的文本信息和各种数据的虚拟

屏幕，达到了增强现实的效果。Google Glass 通过捕捉眼球的运动来获取用户指令^[19]，用户只要眨眨眼就能拍照上传、收发短信、查询天气路况等操作。Google Glass 还配备了音控输入设备，用户可以方便的通过麦克风来启动谷歌眼镜，只要说出“ok，glass”即可。



图 19 Google Glass

而美国康宁 Corning 公司则提出了更具有创新性的未来人机交互构想。在其公司推出的关于未来生活的想象的视频中，无处不在的“玻璃”提供了虚实融合的平台；语音、手势辅助触控操作提供了方便快捷的交互手段。智能玻璃成为了移动终端和交互平台，并与家具用品、交通工具，教学设备等融合成一个智能环境。这种模式提供了一种脱离穿戴设备，依托基础设施实现“虚实融合”的思路。



图 20 A day made of glass

从第六感到谷歌眼镜，再到康宁公司的未来愿景，可以看出，穿戴设备的形态将不断精炼，能力将更为强大。未来的设备隐藏到周边环境中，用户将摆脱设备的束缚，同时得到前所未有的服务和交互体验。

3 国内研究进展

3.1 穿戴式计算国内研究现状

3.1.1 关键技术研究现状

3.1.1.1 传感器与传感能力

国内相关产业起步相对较晚，目前在芯片、电子元件和新材料方面相对落后。国内

的传感器（尤其是高端传感器）严重依赖进口，国产化缺口巨大。根据社科院发布的《2014 年中国经济形势分析与预测》，国内的传感器芯片进口占比高达 90%。

随着穿戴式计算在国内不断兴起，伴随物联网战略的落实，国内 MEMS 传感器公司近两年来取得了较大的进步。如苏州固锝，其于 2011 年通过收购进军 MEMS 传感器开发领域，2014 年已能够研发高质量的三轴加速度传感器、陀螺仪、压力传感器等，并已向苹果、三星、飞利浦、佳能等国际企业供货。目前国内传感器行业通过海外并购或者代工封测两种方式正逐步缩小与世界的差距，并开始生产高精度、符合穿戴式设备需求的传感器。

3.1.1.2 数据分析识别

我国穿戴式硬件设备方面仍然落后于国外，但在支撑软件和服务方面已基本与国际水平同步。如，中国计量学院的潘巨龙和浙江大学的李善平等^[20]在 2007 年即提出过一种基于无线传感器网络的社区保健监测系统，结合人体生理参数的无创连续监测技术和穿戴式医疗仪器的开发，设计了一种适用于慢性病人的监测系统。除了在医疗方面的运用外，国内研究者在 MEMS 类的传感器数据分析和应用上也得到快速发展，如基于压力传感器制作的可检测跌倒的智能鞋^[21]，采用了薄膜式压力传感器，将传感器安置于鞋垫，用于采集人体运动中的脚底压力信息，采用阈值分析与支持向量机算法相结合的方法对脚底压力值进行分析，判断人体是否跌倒。浙江大学杨峰等设计了一种基于智能手表的交互系统 MagicWatch，利用智能手表配备的加速度、陀螺仪等传感器识别用户的各种手势，进而理解用户意图，并在云端系统的支持下，结合环境上下文完成指定任务，如对当前指向物体进行信息提示、设备间信息迁移等。

3.1.1.3 穿戴式计算模型和应用服务

（1）穿戴式计算模型

和国际主流研究情况类似，国内的穿戴式计算模型大多数也是基于穿戴式计算与云计算相结合的方式。得益于国产穿戴式设备较低的售价，国产穿戴设备在我国市场的普及速度远超谷歌眼镜等进口设备。

除了针对单一穿戴式设备进行研究探索，国内对于穿戴式设备间的协同和计算模型也有深入的研究。

浙江大学从物理空间和信息空间融合的背景出发，根据以人为中心的不同实体之间相互作用的基本原理，提出了基于物理场理论的智能影子模型^[22,24]（图 21）。用户在 CyberSpace 中存在一个自己的数据影子，随时随地与 Physical Space 中真实的“我”相互感知与交互，记录人生信号；计算基础设施中存在一个服务影子，形成个性化、移动式的虚拟化智能空间，随时随地为用户提供智能化普适服务。智能影子的计算范式是从对物理世界的“感知”到反作用至物理世界的“交互”这两个操作交替螺旋式反馈上升的过程，将虚拟空间（服务空间）与用户所在物理空间建立动态映射关系，从而达到从人生记录到人生服务，使用户得到个性化、以人为中心的服务。在智能影子模型中，穿戴式设备是随时随地感知用户状态，并提供普适服务的关键载体。

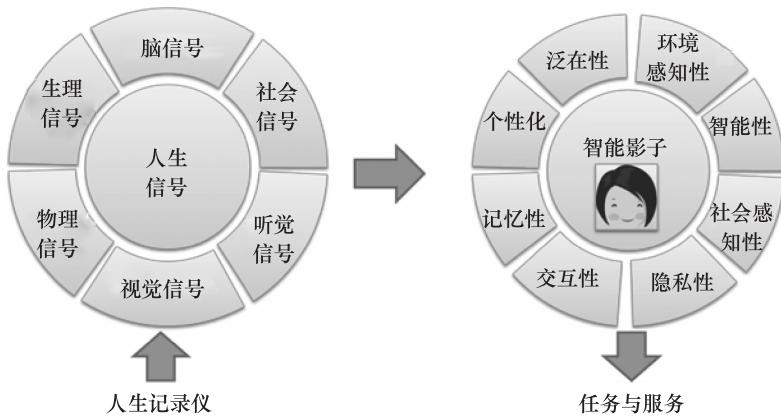


图 21 智能影子用户模型

依托社会计算、众包等理论，西工大郭斌^[23]等研究者尝试基于穿戴设备和其他智能终端构建大规模协作式感知网络，即通过用户随身携带或穿戴的设备以及基础设施中的其他传感器，实时感知识别大量社会个体的行为，并进一步分析挖掘群体社会交互特征和规律。



图 22 群智感知计算体系架构

(2) 穿戴式计算生态体系

在国内工业界，咕咚运动（图 23）是穿戴式计算模型实践较为成功的厂商之一。其

提出了基于硬件、软件和 SNS 的物联网社区 2.0 概念，在硬件上致力于通过各种穿戴式设备采集用户的信息，如咕咚智能手环可以 24 小时记录用户日常活动数据，包括步数、距离、卡路里消耗等，给出运动不足提醒，也能记录分析用户的睡眠质量。在软件上，咕咚的穿戴式设备可以和智能手机或者 PC 进行连接，将相关的数据进行可视化，用户可以方便快捷地了解到自己的运动健康信息等。在云端，用户将自己的运动数据上传，用户之间可以进行互相竞赛、鼓励，从而实现对运动的正向激励。与此同时，开发者可以基于开放的 Codoon API 开发第三方应用。咕咚试图通过这种结合了终端感知、设备协同、云平台和开放 API 的方式建立一个完整的穿戴式生态网络。



图 23 咕咚运动的生态系统

国内的百度等互联网巨头也开始进军穿戴计算领域，并推出了 Baidu Eye 等穿戴式设备。不仅如此，百度公司还将其云平台向穿戴设备厂商开放，通过授权和认证的厂商即可接入百度云，充分利用百度在计算、分发、推广等方面的优势，并与其他接入百度云的穿戴设备实现协同工作。百度试图通过开放接口、数据挖掘、平台支持等方式来建设类似 APP Store 的生态圈，并有意借自己的大数据平台和用户群优势，建立穿戴式设备行业标准。

(3) 面向医疗的穿戴式计算

面向医疗领域的穿戴式计算和云计算的相结合更是大势所趋。首先，穿戴式设备使得连续和大量的数据采集成为可能。设备采集到数据后可以及时上传，历史数据可以积累，趋势可以被掌握，后续的诊断基于大量的数据更准确可靠。其次，连续检测、大数据量分析对于医疗诊断具有重要的意义，如对于过失性心率失常，只有连续的监测才有可能捕捉到，而连续监测血氧则可以针对性地对夜间睡眠呼吸暂停症进行报警。最后，医疗穿戴式计算突破以往的对人群统计数据建模和分析的算法，而是可针对所采集的个人数据进行分析，显著提高服务的个性化程度。

正是因为上述原因，许多医疗信息化厂商开始涉足穿戴计算领域。天津九安医疗 iHealth 针对市场现状，和多个机构合作，推出了一系列的市场化产品。如妊娠糖尿病行为干预系统，其为与北京万康、北京中医药大学的合作项目，针对妊娠期糖尿病患者的

日常行为监护及干预方案推送。该系统通过穿戴式设备监控孕妇的生理信息，并将数据发送到定制的 APP 进行分析，当存在异常状况时，云端会结合医疗机构管理人员的建议，向孕妇推送治疗方案。东软的熙康系统（图 24）则采用一体化健康监测设备，通过互联网，将区域医疗中心、基层卫生服务机构的医疗保健服务与个人、家庭的动态健康管理以及医疗监控与管理部门的数据档案系统进行无缝链接，向用户提供完善的健康服务。此外，国内的中兴物联等企业也推出了类似的结合穿戴式设备的医疗健康服务平台，并已经在部分城市和社区进行了试点应用。



图 24 东软熙康健康服务平台

3.1.2 典型产品和服务

如前文所述，国内的穿戴式产品主要集中在运动和医疗两个方面，在市场化和产业化程度上正逐渐追赶国外。

表 2 国内的商用穿戴式智能设备

设备名称	设备功能
咕咚智能手环	使用蓝牙传输，实时记录日常活动，运动步数、距离以及卡路里，监测睡眠质量，结合百度云提供可视化和数据分析平台
咕咚蓝牙心率带	通过蓝牙传输，实时记录运动心率，实时传输到手机，提供运动状况监测
360 儿童卫士手表	远程定位儿童位置、远程录音、运动计步，距离超出提醒。结合智能云端，记录儿童行进方式和轨迹

(续)

设备名称	设备功能
华为 TalkBand B1	智能手表形态，和蓝牙耳机相结合，来电时作为蓝牙耳机使用，正常状况下可以追踪用户的睡眠和健康状况
BrainLink 智能头箍	基于脑电波传感技术，提供游戏娱乐功能，可以训练儿童的自控力和专注力。与智能手机结合从而实现功能拓展
MUMU 百度云血压计	方便地记录用户血压数据，结合智能手机形成简单明了的趋势示意图，提供健康变化消息提醒。依托百度云平台提供数据存储和可视化

从上表可以看出，与国外成熟的穿戴式设备产业链相比，国内的产品在局部有改善，但总体创新性相对不足，缺乏像谷歌眼镜那样突破性的产品。但无可否认的是，随着诸如华为、百度、360 等移动互联网巨头在产业界的推动（图 25），国内穿戴式设备的发展必将进入快车道，未来可期。



图 25 整合百度云的咕咚手环和 360 儿童卫士

3.2 面向穿戴计算的新型交互技术

新型的移动智能终端和多样的交互设备的涌现，为多种感知和交互技术的融合提供了可能。国内的研究者在这方面提出了许多交互模型，同时交互技术的应用也得到快速发展。

3.2.1 交互模型

(1) 多模协同的体感互动

传统基于单一种类传感器和单一形式信号的体感互动系统已不再能满足当前的交互需求。当前的体感互动系统涉及的传感器种类较多，采集的数据包括生理参数数据、肢体运动数据和视觉捕获数据等，具有异构显著、尺度各异的特点。常规的单通道传感信号数据特征提取方法对体感互动表达的区分性和鲁棒性远远不够。同时各模特征以简单的叠加方式融合成单一的特征向量时，它们之间可能存在相互补充或互为冗余的信息。

电子科技大学移动计算研究中心提出了多模信号协同的体感互动模型（图 26），包括在单模数据集上的特征提取方法的基础上，侧重于挖掘对于体感互动表现最好的特征组合，并以此提取鲁棒性强的新型语义特征表达，对各模特征进行最大化的压缩组合以及最优化的协同提取，对人体感知数据在特征层和决策层进行多模识别融合，进而对用户意图进行识别与理解。

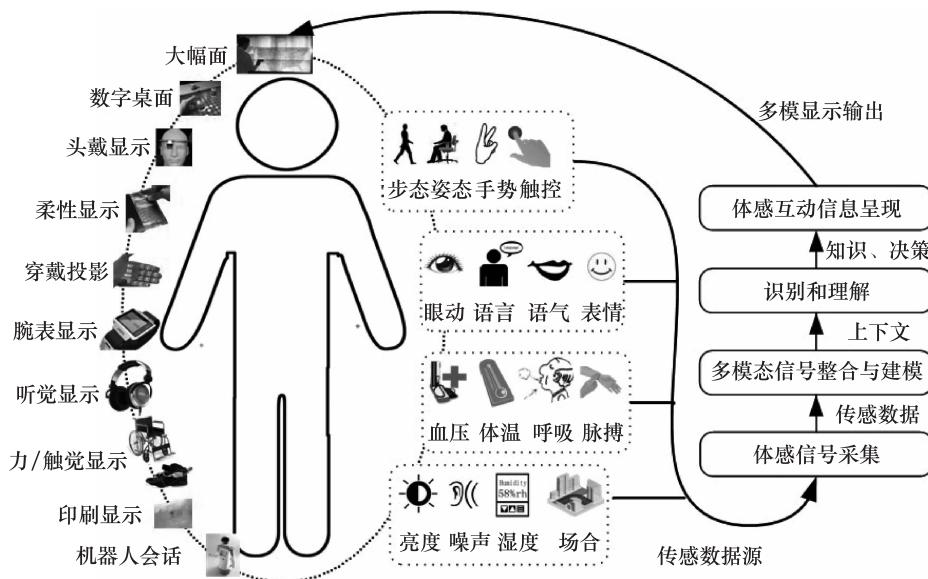


图 26 多模协调体感互动框架

(2) “虚拟感官”

由于传统的人机交互需要用户主导完成，且系统只对规则内的输入进行响应，这无疑破坏了用户体验的灵活性，束缚了用户使用方式。如何解放用户，使用户实现更自然地人机交互成为当前研究热点。

北京科技大学人工生命与智能软件实验室提出了“虚拟感官”的拟人化的交互系统理念，“虚拟感官”是将计算机系统拟人化，输入设备作为其“感官”，让计算机系统来“感受”并理解人的行为意图，主动对人的行为做出反映，这将很好的解决计算机系统只响应用户规则内输入的问题，从而提升系统的用户体验。虚拟感官实现自然人机交互示意图见图 27。

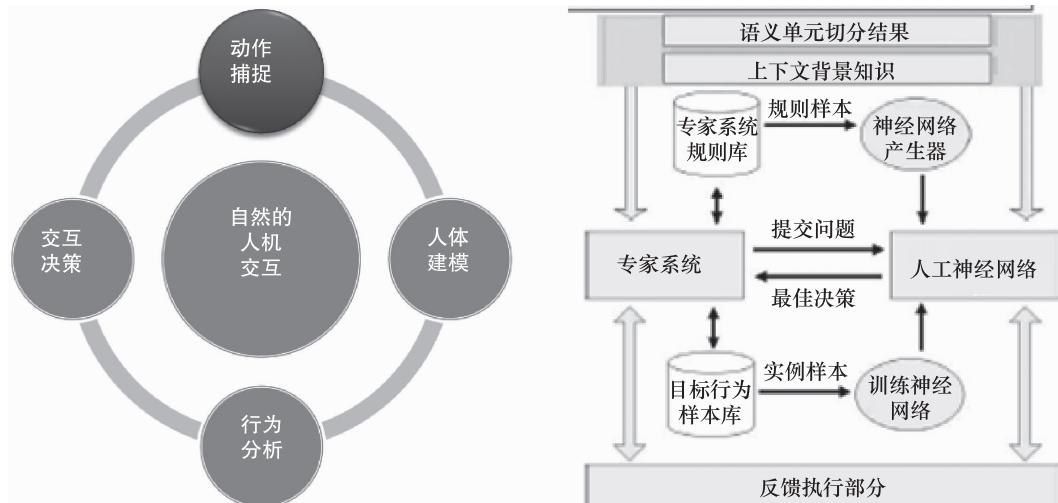


图 27 虚拟感官实现自然人机交互

通过改变传统的“输入 – 规则 – 反馈”模式，让计算机主动检测用户动作和行为，继而提供相应的服务，构建“感知 – 知识 – 服务”模式，解放系统在交互中对输入方式的束缚。

通过与情感计算技术相结合，具有“虚拟感官”的系统通过建立“心智模型”，主动理解用户，取代人对系统的理解，可以使人通过本能的动作来操控系统。同时，由于系统的不断“学习”，对用户的行为有一定适应和学习能力，可根据当前环境和用户状态，主动为用户提供服务，达到真正的自然交互。

3.2.2 交互应用

(1) 手语识别

手语识别是解决聋哑人与正常人沟通的主要技术手段，其难点主要包括手语运动数据获取不完整、多维手语运动数据识别、大词汇量识别、非特定人的识别、手语多模态表达的复杂性及手语运动数据的重定向等。

中国科学院计算技术研究所使用新型交互技术在辅助聋哑人与正常人交流的应用上取得了突出成果。手语识别系统主要分为基于数据手套的识别和基于视觉（图像）的手语识别系统（图 28）。



图 28 基于数据手套的手语识别系统与基于视觉的手语识别系统

基于数据手套的手语识别系统，是利用数据手套和位置跟踪器测量手势在空间运动的轨迹和时序信息，主要用于数据采集；基于视觉的手语识别系统则通过计算机三维图像采集设备获得聋哑人的手语数据，采用模式识别算法，结合上下文知识，获知手语含义，进而翻译成语音，传达给不懂手语的正常人。

(2) 远程沉浸式交互

目前常用的远程交互方式主要有文字交互、电话和远程视频等。由于电话只能呈现远程用户的声音，而现有的视频对话系统则存在较严重的空间隔离感，而且操作较为复杂，交流介质较为单一，在情感交流方面存在严重不足。中科院普适计算研究中心研制的爱心小屋——基于人机物三元融合端计算技术的远程亲情互动平台研制了一个低成本、沉浸式、易操作、高保真的远程自然交互平台（图 29）。



图 29 爱心小屋远程亲情互动平台

该项目无需穿戴设备即能提供沉浸式的交互体验。其通过音视频增强处理技术，实现“面对面”的沉浸式视频；通过远程协同交互技术，实现“手把手”的直觉式协作；通过人、机、物三元交互技术，实现“心连心”的融入式交互。其中包含多项核心技术，例如启发式手势操控、精准人像提取、基于精准对象分割的虚实融合视频合成技术、基于深度伪 3D 的自然眼神交流等。

4 国内外研究进展比较

国外由于较早研究穿戴式设备，其研发水平起点高于我国，但是得益于中国的庞大市场和数量庞大的研究团队，我国正在穿戴式计算领域奋起直追。根据 Frost & Sullivan 咨询公司发布的《2013 年中国智能穿戴设备市场研究报告》，仅 2012 年的国内穿戴式设备市场规模已经达到 8.9 亿元，预计到 2015 年，市场规模将达到 26.1 亿元。从报告中可明显看出，随着国家战略倾斜，在工业界和学术界大力推动下，我国穿戴式计算增速迅猛。对比国内外发展现状，在学术领域和工业化领域体现出如下的特点：

(1) 学术领域

穿戴式计算概念源自国外，研究人员来自各个高校、研究所等学术机构，也包括一些国际工业巨头的实验室和小型科技公司。工业界研究者的积极参与，使得穿戴式计算的研究工作目标和导向更为清晰。而国内研究人员主要来自专业学术机构，学科交叉不够充分，同时缺乏市场导向和产业化激励，成果转化路径不够系统，成果产业化程度不高。另外，由于穿戴式计算的研究涵盖了计算机科学、材料科学、人机工程、通信工程、医疗、生物等多种领域，国内研究由于这些领域的短板也造成了整体上研究进展的相对落后。

然而，得益于穿戴式计算在国内的庞大潜力，硬件层次上，国内的研究机构尤其是各个大学的团队正与科技产业公司形成越发紧密的合作，一部分自主传感器研究领域已

有一定的突破，如中科院深圳先进技术研究院已能开发完全自主知识产权的医学集成电路芯片。软件层次上，国内在穿戴式计算服务模型、穿戴式计算数据分析等方面的研究也取得了可喜的进展，部分成果受到国内外研究者的好评和引用。但总体而言，国内的学术和产业界研发工作任重道远，在跟随国外研究进展的同时，还需力争在理念和产品上引领创新浪潮。

（2）工业化领域

国外工业界在半导体、芯片领域明显领先于国内，各种规格、精度、尺寸的传感器均有成熟的解决方案和实际产品，包括 MEMS 微机电系统传感器、监测血糖血压血氧的医疗传感器、高精度微小体积的温湿度传感器、压力传感器和加速度传感器等。国内的传感器硬件领域基础仍比较薄弱，大量的精密传感器需要通过进口才能满足需求。但是近年来国内的传感器厂商通过并购、自主研发，已经能够相对独立地生产符合穿戴式计算要求的传感器，包括微机电系统传感器、各类压力传感器等，且已经获得了国际知名智能设备厂商的认可与采购。

在商用产品方面，国际厂商推出了诸如 Google Glass、Meta 智能眼镜等极富创意和市场前景的穿戴式设备。国内厂商针对国内的消费水平和市场开发状况，致力于智能手环、智能鞋以及智能医疗设备等领域取得了良好的进展，诸如咕咚手环等产品已经在国内有了一定的知名度，其生态系统也正逐步建立；百度等厂商正在大力推动开放的穿戴设备云端协作平台，并据此建立其穿戴设备生态系统。然而总体来看，受限于技术和设计理念的局限，我国的穿戴计算产品在创新性和技术含量方面还是逊色于国外的产品，但这一情况正在迅速改善。

5 发展趋势与展望

从穿戴式计算的概念被提出到今天，穿戴式计算的发展已经历了 40 多年。早期的设备功能单一，大多是离线工作，能提供的信息和服务能力极其有限。近年来，得益于物联网、云计算等技术浪潮的推动，依托 MEMS、材料、数据分析和交互技术的发展，各类新型、异构的穿戴设备从云端得到强大支持，并通过设备协同提供更为复杂的服务。

随着这些技术浪潮的进一步发展，我们认为，穿戴设备未来最明显的发展趋势是不断小型化、微型化、异构化，其传感、显示和交互能力日益依赖于基础设施的支持，并最终嵌入直至“消失”在基础设施中，使用户彻底摆脱设备的束缚。

在这一趋势下，还衍生出了一系列挑战性问题有待突破，例如：

- 1) 目前的穿戴设备在“解放双手”的同时，给身体的其他部位带来了负担。需要怎样的穿戴设备来进一步解放用户，让用户无障碍地融入信息环境？
- 2) 如何在不同环境中记录、理解用户的行为，并基于这些理解，跨越用户学习和适应穿戴设备的过程，实现无障碍、自由的交人机交互？

3) 在日益微型化的穿戴设备上, 用户无法像在PC上那样便利地进行信息筛选, 因此对信息的精准性等质量要求极高。如何提供高质量信息, 实现新型穿戴终端“所见即所需”的要求?

4) 如何构建连续、普遍存在的计算/交互基础设施, 增强穿戴设备功能, 弱化穿戴设备对本身硬件能力的依赖, 使其不断小型、微型化, 直至“消失”?

总之, 穿戴式设备集中体现了多学科的研究成果, 其独有优势使其不断强化甚至取代PC及智能手机的功能。穿戴式计算极充分地体现了“围绕人的服务”、“消失的计算”等特征, 是普适计算的重要载体和平台, 衍生出一系列新型计算范式、交互方法, 并带来数据分析、服务提供、平台构建等一系列理论问题。同时, 穿戴计算可能演化出未来的新型商业模式, 是智能手机之后的全新技术换代和升级。穿戴式计算浪潮将使学术界和工业界迎来一轮新的发展契机。

参考文献

- [1] Mann S. Wearable computing: A first step toward personal imaging[J]. Computer, 1997, 30(2): 25-32.
- [2] 王启明. 一场穿戴式技术革命正在北美地区兴起[J]. 全球科技经济瞭望, 2013(10).
- [3] Bryzek J. Impact of MEMS technology on society[J]. Sensors and Actuators A: Physical, 1996, 56(1): 1-9.
- [4] 滕晓菲, 张元亭. 移动医疗: 穿戴式医疗仪器的发展趋势[J]. 中国医疗器械杂志, 2006, 30(5): 330-340.
- [5] Lee Y D, Chung W Y. Wireless sensor network based wearable smart shirt for ubiquitous health and activity monitoring[J]. Sensors and Actuators B: Chemical, 2009, 140(2): 390-395.
- [6] 杨峰, 李石坚. 穿戴式计算[J]. 中国计算机学会通讯, 2014, 1(10): 76-81.
- [7] Mistry P, Sadi S, Yao L, et al. User Gesture Input to Wearable Electronic Device Involving Outward-Facing Sensor of Device: U. S. Patent Application 14/015, 909[P]. 2013-8-30.
- [8] 龙昭华, 李景中, 蒋贵全, 张林. 基于无线传感器网络的普适计算研究[C]. 第六届和谐人机环境联合学术会议(HHME2010)、第19届全国多媒体学术会议(NCMT2010)、第6届全国人机交互学术会议(CHCI2010)、第5届全国普适计算学术会议(PCC2010), 2010.
- [9] 王宏安, 田丰, 翟树民. 人机交互[J]. 中国计算机学会通讯, 2011, 7(11).
- [10] 滕继濮. 人机交互: 技术为人服务[N]. 科技日报, 2011.
- [11] Wu J, Pan G, Zhang D, et al. MagicPhone: pointing & interacting[C]//Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing- Adjunct. ACM, 2010: 451-452.
- [12] 袁彬, 肖波, 侯玉华, 等. 移动智能终端语音交互技术现状及发展趋势[J]. 信息通信技术, 2014(2).
- [13] Argall B D, Billard A G. A survey of tactile human-robot interactions[J]. Robotics and Autonomous Systems, 2010, 58(10): 1159-1176.
- [14] Kim D H, Lu N, Ma R, et al. Epidermal electronics[J]. Science, 2011, 333(6044): 838-843.
- [15] Fishel J A, Loeb G E. Sensing tactile microvibrations with the BioTac—Comparison with human sensitivity [C]//Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International

- Conference on. IEEE, 2012: 1122-1127.
- [16] 刘大有, 刘春辰, 王生生. 环境智能中上下文推理方法研究综述[J]. 模式识别与人工智能, 2011(5).
- [17] Scherer K R, Bänziger T. Emotional expression in prosody: a review and an agenda for future research [C]. Speech Prosody 2004, International Conference, 2004.
- [18] Mistry P, Maes P. SixthSense: a wearable gestural interface[C]. ACM SIGGRAPH ASIA 2009 Sketches, 2009.
- [19] Noris B, Keller J, Billard A. A wearable gaze tracking system for children in unconstrained environments [J]. Computer Vision and Image Understanding, 2011, 115(4).
- [20] 潘巨龙, 李善平, 吴震东. 基于无线传感器网络的社区保健监测系统[J]. 中国计量学院学报, 2007, 18(2): 136-140.
- [21] 石欣, 张涛. 一种可穿戴式跌倒检测装置设计[J]. 仪器仪表学报, 2012, 33(3): 575-580.
- [22] 潘纲, 张犁, 李石坚, 吴朝晖. 智能影子(SmartShadow):一个普适计算模型[J]. Journal of Software, 2009, 20: 40-50.
- [23] 郭斌, 张大庆, 於志文, 周兴社. 数字脚印与“社群智能 c”[J]. 中国计算机协会通讯, 2011, 7 (3): 53-60.
- [24] Pan G, Zhang L, Wu Z, et al. Pervasive Service Bus: Smart SOA Infrastructure for Ambient Intelligence [J]. 2012.

作者简介

李石坚 博士, 浙江大学计算机学院副教授, 主要研究方向为普适计算、人工智能。E-mail: shijianli@zju.edu.cn。



班晓娟 博士, 北京科技大学计算机与通信工程学院教授, 主要研究方向为自然人机交互、普适计算等。E-mail: banxj@ustb.edu.cn。



叶振宇 博士研究生, 浙江大学计算机学院, 主要研究方向为普适计算。E-mail: enyzy@zju.edu.cn。



沈 晴 博士研究生，北京科技大学计算机与通信工程学院，主要研究方向为人机交互、计算机视觉、机器学习。E-mail：shenqingcc222333@gmail.com。



潘 纲 博士，浙江大学计算机学院教授，主要研究方向为普适计算、计算机视觉、智能系统等。E-mail：gpan@zju.edu.cn。



关键词索引

- 表示学习 119,121,129,129
穿戴式计算 350,350,350,350,351,351,351,351,351,352,355,357,358,358,360,364,365,365,365,365,365,366,366,367,367,367,372,372,372,372,372,372,373,373,373,373,374,374,374
大规模在线开放课程 218
大数据分析 189,21,218,218,22,221,221,221,248,248,248,248,248,249,249,250,251,4
分层架构 119,119,130
个性化服务 218,218,221,222,233,242,242,245,250,257
互联网体系结构 67,67,67,68,68,79,96,111,111,111,111,113,113,113,114,115,116,116,116,116,116,118,118
计算机辅助设计 138,188,188,188,188,188,188,188,188,189,208,208,216,216
计算机图形学 188,189,190,190,190,190,191,193,194,205,207,216,216,216,217,217,333
健康促进 296,301,302,302,302,302,302,302,302,302,302,302,302,302,303,303,303,303,303,303,303,303,305,305,305,305,306,306,308,309,309,309,312,313
健康感知 296,296,296,296,296,297,297,297,298,298,298,298,298,298,298,298,298,298,298,298,299,299,299,299,299,300,303,303,304,304,304,306,306,306,306,306,306,307,308,308,308,308,308,308,308,308,309,309,309,309,310,310,310
健康计算 296,296,300,300,300,308,308,314,314,314
健康评估 296,297,301,302,304,305,307,307,308,309,309,309,309,309,309,309,313,313
卷积神经网络 119,120,123,123,123,128,128,128,129,129,129,129,206
可视化与可视分析 188,188,189,192,200,204,206,208
目标跟踪 315,320,322,322,323,323,323,323,324,333,336,339
目标检测 128,129,137,315,315,316,320,320,320,320,320,320,320,321,321,321,321,321,321,321,321,322,322,332,332,332,333,333,333,333,333,333,333,335,336,336,336,336,336,336,336,336,336,339,339,342,347,349
内存计算 1,1,1,1,2,2,2,2,4,7,11,11,11,12,12,12,12,12,12,13,13,14,14,14,15,15,19,19,19,19,20,20,20,20,20,24,24,26,27,27,27,27
内容中心网络 67,68,79,115

- 127, 127, 128, 128, 128, 128, 128, 128, 128, 129, 129, 129, 129, 129, 129, 130, 130, 130, 131, 131, 131, 131, 136, 137, 245, 250, 341, 341, 341
- 视频监控 315, 315, 315, 315, 315, 316, 316, 316, 317, 317, 317, 317, 319, 320, 320, 324, 325, 326, 327, 327, 329, 329, 329, 329, 330, 330, 330, 330, 332, 332, 332, 332, 333, 333, 333, 334, 334, 336, 336, 336, 336, 337, 341, 341, 342, 342
- 视频增强 315, 315, 316, 323, 324, 333, 333, 336, 336, 336, 337, 337, 337, 337, 339, 347, 372
- 受限玻尔兹曼机 119, 121, 124, 124
- 数据密集计算 1, 1, 1, 1, 2, 2, 2, 2, 20, 20, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27
- 数据挖掘 3, 3, 3, 4, 7, 10, 14, 27, 193, 206, 206, 223, 227, 229, 229, 257, 258, 280, 280, 294, 294, 294, 301, 301, 301, 309, 313, 367
- 数据中心网络 35, 35, 35, 35, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 36, 37, 37, 37, 37, 38, 38, 38, 39, 39, 39, 39, 39, 39, 40, 40, 40, 40, 41, 41, 41, 41, 41, 41, 41, 41, 42, 42, 42, 42, 42, 42, 42, 42, 42, 42, 43, 43, 43, 43, 43, 43, 43, 44, 44, 44, 44, 44, 44, 45, 45, 45, 48, 48, 48, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 49, 50, 50, 50, 50, 50, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 51, 52, 52, 52, 52, 52, 52, 52, 52, 52, 53, 53, 53, 53, 54, 54, 54, 54, 54, 54, 54, 62, 64, 64, 65, 65, 65, 113, 113, 113, 113
- 数字媒体 188, 188, 189, 189, 189, 189, 189, 191, 191, 191, 191, 192, 192, 192, 192, 199, 199, 199, 199, 200, 200, 204, 204, 204, 204, 204, 206, 206, 206, 206, 206, 208, 315, 332, 348, 349
- 图计算 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10, 10, 11, 11, 11, 11, 26, 27, 29
- 图形绘制 188, 188, 188, 189, 189, 189, 189, 189, 189, 190, 190, 190, 191, 196, 196, 203, 203, 204, 205, 206, 206, 206, 208
- 微课程 218, 227, 234, 234, 237, 237, 240, 240, 240, 240, 242, 242, 244, 244, 244, 244, 245, 245, 245
- 未来互联网 50, 67, 67, 68, 74, 85, 96, 104, 108, 108, 108, 109, 109, 109, 109, 111, 112, 112, 113, 114, 115, 115, 115, 118, 118
- 系统软件 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 11, 11, 12, 12, 19, 19, 19, 20, 20, 27, 27, 27, 33, 34, 34
- 信息检索 258, 268, 273, 278, 294
- 信息中心网络 67, 114
- 行为识别 315, 315, 316, 326, 326, 328, 328, 328, 328, 334, 337, 341, 341, 341, 347, 347, 348, 355, 360
- 虚拟网络 , 35, 43, 43, 44, 44, 47, 51, 51, 53, 53, 54, 101, 103, 103, 107, 108

自编码器 119, 123, 126, 126, 126, 126, 126, 128, 129

作者索引

- 班晓娟 350
班晓娟 375
鲍虎军 188
毕 军 111, 112, 67
陈贵海 35, 66
陈海波 1, 2, 2, 2, 2, 29, 33
陈 为 188, 208, 216
杜海鹏 218
封举富 119, 138
冯结青 188, 208, 216
冯 时 258, 294
傅慧源 315, 347, 348
巩敦卫 139, 166, 184, 184, 184, 184, 184, 185, 187
郭得科 35, 65
何 源 35, 65
胡宇翔 118, 67
胡占义 119, 138
黄铁军 315, 347, 348
江 贺 139, 187
姜安琦 258, 295
姜育刚 315, 348, 348
李 波 293, 315, 347, 348
李 丹 35, 60, 64
李瑞轩 258, 294
李石坚 350, 374, 375, 375
李玉华 258, 294
李振宇 118, 67
李 征 139, 187
廖小飞 1, 2, 2, 32, 34
林 强 296, 314
刘方明 35, 60, 65
刘利刚 188, 202, 205, 208, 216
罗洪斌 117, 67
罗英伟 1, 2, 2, 34
马华东 315, 347, 348
倪红波 296, 314
聂长海 139, 184, 185, 185, 185, 187
潘 纲 350, 375, 376
申德荣 294
沈 晴 350, 376
汤 庸 258, 294
田 锋 218, 234, 257
王大玲 258, 293
王立威 119, 138
王 锐 188, 196, 197, 197, 197, 203, 208, 217
王天本 296, 314
王 柱 296, 314
魏笔凡 218, 257
邢春晓 258, 294
薛向阳 315, 348, 348
叶振宇 350, 375
于 戈 258, 29, 29, 293
于俊清 315, 348, 349
于 旭 258, 295
张松海 188, 208
张松海 217
张未展 218, 254, 257
郑 锦 315, 347, 349
郑庆华 218, 254, 257
周兴社 296, 304, 314, 375