



中国计算机学会文集

China Computer
Federation Proceedings
CCFP 0025

CCF 2014-2015中国计算机科学技术 发展报告

中国计算机学会 主编



机械工业出版社
China Machine Press

图书在版编目 (CIP) 数据

CCF 2014-2015 中国计算机科学技术发展报告 / 中国计算机学会主编. —北京: 机械工业出版社, 2015.10
(中国计算机学会文集)

ISBN 978-7-111-51829-7

I. C… II. 中… III. 计算机科学—发展—研究报告—中国—2014 ~ 2015 IV. TP3-12

中国版本图书馆 CIP 数据核字 (2015) 第 239802 号

本书是由中国计算机学会学术工作委员会组织编写的具有权威性的计算机科学技术年度发展报告，总结了 2014—2015 年我国计算机科学技术发展的部分重要成果，选定了天地一体化网络研究进展与趋势、“互联网 +”战略的研究进展与趋势、基础计算系统可靠性研究进展、网络信息安全科技与应用发展综述、大数据机器学习的研究进展与趋势、新型数据管理系统研究进展与趋势、经验软件工程的挑战和发展趋势、知识型服务计算、深度学习与媒体计算、文本自动生成研究进展、工业控制计算机研究进展与发展趋势、抗恶劣环境计算机技术的现状及发展趋势 12 个方向进行总结。书中报告分别由活跃在这些研究方向上的一线科研人员撰写，详细介绍了相应研究方向在研究、开发和应用等方面取得的进展，并对国内外在该研究方向上的研究现状进行了对比，分析了未来可能的发展趋势。

本年度报告是广大计算机科学技术人员了解当前计算机科学技术发展动态的一个渠道，也适合本领域决策人员和科研人员参考。

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：张梦玲 王春华

责任校对：董纪丽

印 刷：中国电影出版社印刷厂

版 次：2015 年 10 月第 1 版第 1 次印刷

开 本：185mm×260mm 1/16

印 张：25

书 号：ISBN 978-7-111-51829-7

定 价：99.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：(010) 88378991 88361066

投稿热线：(010) 88379604

购书热线：(010) 68326294 88379649 68995259

读者信箱：hzjsj@hzbook.com

版权所有 • 侵权必究

封底无防伪标均为盗版

本书法律顾问：北京大成律师事务所 韩光 / 邹晓东

前　　言

计算机科学技术的高速发展为国民经济发展提供了强大的推动力，特别在“互联网+”时代，计算机科学技术促进了传统产业的革新和现代服务业的兴起，越来越深刻地影响着社会经济生活的运行。

中国计算机学会（CCF）编写的中国计算机科学技术年度发展报告，旨在记录计算机科学技术各领域的发展现状和热点问题，展望未来的发展趋势和动向，为科研人员、计算机领域教育认识、在学研究生、企业提供宏观技术参考。

本年度报告的形成经过了广泛征稿、严格评审、组织学科发展研讨会、统一终审等过程，最后从泛在网络、系统可靠性、安全/软件工程和智能技术/应用等4个方面，选定了天地一体化网络研究进展与趋势、“互联网+”战略的研究进展与趋势、基础计算系统可靠性研究进展、网络信息安全科技与应用发展综述、大数据机器学习的研究进展与趋势、新型数据管理系统研究进展与趋势、经验软件工程的挑战和发展趋势、知识型服务计算、深度学习与媒体计算、文本自动生成研究进展、工业控制计算机研究进展与发展趋势、抗恶劣环境计算机技术的现状及发展趋势12个方向进行总结。

本年度报告总结了2014—2015年我国计算机科学技术发展的部分重要成果，分别由活跃在这些研究方向上的一线科研人员撰写，详细介绍了相应研究方向在研究、开发和应用等方面取得的进展，并对国内外在该研究方向上的研究现状进行了对比，分析了未来可能的发展趋势。从一定角度反映了我国计算机科学和技术工作当前的研究进展和现状，对学术研究和人才培养有重要参考价值，对促进我国计算机科学技术的发展、推动我国信息化进程也会起到重要作用。本年度报告是广大计算机科学技术人员了解当前计算机科学技术发展动态的一个渠道，也适合本领域决策人员和科研人员参考。

由于策划、组稿时间短，报告在形式和内容安排上都可能存在不妥之处，希望广大读者对本报告的组织、编写工作多提宝贵意见和建议，以便不断提高年度报告的质量，提升其参考价值。报告中的观点仅代表撰稿人的个人意见。

最后，谨对为本年度报告贡献稿件的所有专家表示感谢。中国计算机学会学术工作委员会的委员们在本年度报告评审、内容研讨和综审过程中付出了辛勤的劳动，特别是学术工委秘书、北京大学刘謙哲副教授在稿件收集整理、送审和安排学科发展研讨会的过程中付出了大量的辛勤劳动；机械工业出版社华章公司的编辑们负责稿件的整理和编辑工作。另外，中国计算机学会郑纬民理事长、杜子德秘书长等对本年度报告的出版也给予了许多指导和具体的支持。在此一并向他们表示感谢。

金芝
中国计算机学会学术工作委员会主任

目 录

前言

天地一体化网络研究进展与趋势 CCF 互联网专委会

1 引言	1
2 我国天地一体化网络的发展愿景	2
2.1 系统能力	2
2.2 典型应用	2
3 国内外现状及差距	3
3.1 国外发展现状	3
3.2 国内发展现状	16
3.3 国内外差距	19
4 体系结构	20
4.1 体系结构的发展现状	20
4.2 体系结构的研究难点	24
4.3 体系结构的发展趋势	25
5 组网架构	25
5.1 网络结构的发展现状	25
5.2 组网架构的研究难点	27
5.3 组网架构的发展趋势	28
6 关键技术	28
6.1 路由技术	28
6.2 传输技术	29
6.3 星上信息处理技术	31
7 结束语	32
参考文献	33
作者简介	35

“互联网 +” 战略的研究进展与趋势 CCF Petri 网专委会

1 发展背景	38
2 国际研究现状	40
2.1 互联网行业发展概述	40

2.2 各国“互联网+”相关战略计划概述与分析.....	41
2.3 “互联网+”实施关键技术发展现状.....	42
3 国内研究现状	50
3.1 我国互联网行业发展概述	50
3.2 “互联网+”战略.....	50
3.3 典型行业“互联网+”应用现状.....	52
4 “互联网+”战略发展思路与对策	52
4.1 数据资源层面的“互联网+”	52
4.2 系统层面的“互联网+”	53
4.3 面向全产业链的“互联网+”	53
5 “互联网+”研究进展与展望	54
5.1 互联网交通	55
5.2 互联网金融	57
5.3 互联网医疗	60
6 结束语	62
参考文献	63
作者简介	65

基础计算系统可靠性研究进展 容错计算专委会

1 引言	66
2 国际研究现状	70
2.1 考虑多故障模式的云计算平台可靠性评估国际研究现状	70
2.2 基于故障注入的云计算平台可靠性评测国际研究现状	72
2.3 云计算平台的开源软件可靠性评估研究国际进展	73
2.4 系统部件维护策略国际研究进展	74
2.5 互联网络可靠性评估国际研究进展	76
3 国内研究进展	78
3.1 基于 MDD 的全系统建模和分析国内研究进展	78
3.2 基于故障注入的云计算平台可靠性评测国内研究进展	79
3.3 构件软件可靠性过程国内研究进展	82
3.4 基于 BDD 的互联网络可靠性国内研究进展	83
3.5 互联网络可靠性参数及优化国内研究进展	85
4 国内外研究进展比较	87
4.1 计算平台和互联网络可靠性评估研究进展比较	87
4.2 基于故障注入的基础计算系统可靠性评测研究进展比较	88
4.3 构件软件可靠性过程研究进展比较	88

4.4 互连网络可靠性参数的可实现性问题、性质和优化问题研究进展比较.....	89
5 发展趋势和展望	89
5.1 多故障模式云计算平台可靠性研究展望	89
5.2 硬件组件维护策略研究展望	90
5.3 互连网络可靠性评估研究展望	91
6 结束语	92
参考文献	92
作者简介	96

网络信息安全科技与应用发展综述 CCF 计算机安全专委会

1 引言	98
2 国外网络信息安全发展现状	99
2.1 各国网络信息安全动态	99
2.2 国外网络信息安全大事件	106
3 国内网络信息安全发展现状	108
3.1 国内网络信息安全动态	108
3.2 国内网络信息安全大事件	112
4 新技术新应用的发展	113
4.1 智慧城市的发展	113
4.2 互联网金融的发展	114
4.3 社交网络的发展	115
5 对策和建议	116
5.1 智慧城市安全对策	116
5.2 互联网金融安全对策	117
5.3 社交网络安全对策	119
6 结束语	120
参考文献	120
作者简介	121

大数据机器学习的研究进展与趋势 CCF 人工智能与模式识别专委会

1 引言	122
2 国际研究进展	123
2.1 数据消减	123
2.2 算法伸缩	127
3 国内研究进展	132
4 发展趋势与展望	134

5 结束语	135
致谢	135
参考文献	135
作者简介	143

新型数据管理系统研究进展与趋势 数据库专委会

1 概述	144
1.1 引言	144
1.2 新型数据管理系统的处理架构	145
1.3 总结	149
2 图数据管理系统	149
2.1 引言	149
2.2 大图数据管理面临的挑战	150
2.3 大图数据管理中的主要研究问题	151
2.4 国外研究现状	154
2.5 国内研究现状	155
2.6 国内外研究进展比较	157
2.7 总结	158
3 流数据管理系统	158
3.1 引言	158
3.2 国际研究现状	159
3.3 国内研究现状	165
3.4 总结	165
4 众包数据管理系统	166
4.1 引言	166
4.2 众包数据管理	168
4.3 国外研究现状	171
4.4 国内研究现状	173
4.5 总结	173
5 在线数据分析与管理系统	174
5.1 引言	174
5.2 主要研究问题	175
5.3 国外研究现状	176
5.4 国内研究现状	179
5.5 国内外研究进展比较	181
5.6 总结	181

6 商业数据管理系统	182
6.1 引言	182
6.2 传统关系型数据库	182
6.3 NoSQL	183
6.4 NewSQL	185
6.5 云关系数据库	186
6.6 NoSQL 的云数据库	187
6.7 总结	188
7 发展趋势	189
7.1 数据融合与知识融合	189
7.2 基于新型硬件的大数据管理	190
7.3 大数据隐私管理	191
8 结束语	192
致谢	192
参考文献	192
作者简介	201

经验软件工程的挑战和发展趋势 CCF 软件工程专委会

1 引言	203
2 经验软件工程的基本研究方法	206
2.1 受控实验	206
2.2 调研	207
2.3 案例研究	208
2.4 经验数据的处理和分析方法	208
2.5 经验分析的质量与实验有效性验证	209
3 软件工程各领域的主要经验研究成果	211
3.1 估算与预测	211
3.2 软件需求、软件结构设计及度量	218
3.3 软件测试与质量保障	222
3.4 软件开发与过程改进	226
3.5 开源与分布开发	229
4 国内外研究工作比较和展望	234
5 结束语	235
致谢	236
参考文献	237
作者简介	251

知识型服务计算	CCF 服务计算专委会
1 引言	253
2 国际研究现状	255
2.1 基于知识的服务发现与推荐	255
2.2 基于知识的服务组合	256
2.3 基于知识的业务流程管理	258
2.4 基于知识的服务质量管理与预测	261
3 国内研究进展	262
3.1 基于知识的服务发现与推荐	262
3.2 基于知识的服务组合	263
3.3 基于知识的业务流程管理	265
3.4 基于知识的服务质量管理与预测	266
4 国内外研究进展比较	266
4.1 基于知识的服务发现与推荐	266
4.2 基于知识的服务组合	267
4.3 基于知识的业务流程管理	267
4.4 基于知识的服务质量管理与预测	268
5 发展趋势与展望	268
5.1 基于知识的服务发现与推荐	268
5.2 基于知识的服务组合	268
5.3 基于知识的业务流程管理	269
5.4 基于知识的服务质量管理与预测	269
6 结束语	270
致谢	270
参考文献	270
作者简介	280
深度学习与媒体计算	CCF 多媒体专委会
1 引言	282
2 国际研究现状	284
2.1 图像检索排序与标注	284
2.2 多模态检索与语义理解	285
2.3 视频分析与理解	287
3 国内研究进展	289
3.1 图像检索排序与标注	289

3.2 多模态检索与语义理解	290
3.3 视频分析与理解	291
4 国内外研究比较	291
4.1 工业应用	291
4.2 算法研究	292
4.3 开源软件	292
5 未来挑战和展望	292
致谢	293
参考文献	293
作者简介	296

文本自动生成研究进展与趋势 CCF 中文信息技术专委会

1 引言	298
2 文本到文本的生成	299
2.1 国际研究现状	299
2.2 国内研究现状	304
2.3 发展趋势与展望	305
3 意义到文本的生成	305
3.1 国际研究现状	305
3.2 国内研究现状	307
3.3 发展趋势与展望	308
4 数据到文本的生成	308
4.1 国际研究现状	308
4.2 国内研究现状	311
4.3 发展趋势与展望	312
5 图像到文本的生成	312
5.1 国际研究现状	312
5.2 国内研究现状	315
5.3 发展趋势与展望	315
6 总结与展望	316
参考文献	316
作者简介	322

工业控制计算机研究进展与发展趋势 工业控制计算机专委会

1 引言	324
2 工业控制计算机	325

2.1 概述	325
2.2 国际研究现状	326
2.3 国内研究进展	335
2.4 国内外研究进展比较	338
2.5 发展趋势与展望	339
3 嵌入式系统	339
3.1 概述	339
3.2 国际研究现状	340
3.3 国内研究进展	347
3.4 国内外研究进展比较	349
3.5 发展趋势与展望	350
4 高可靠容错计算机关键技术	350
4.1 概述	350
4.2 国际研究现状	351
4.3 国内研究进展	355
4.4 国内外研究进展比较	357
4.5 发展趋势与展望	357
5 结束语	358
参考文献	358
作者简介	360

抗恶劣环境计算机技术的现状与发展趋势 CCF 抗恶劣环境计算机专委会

1 引言	361
2 国际研究现状	363
2.1 容错体系结构	363
2.2 软件可靠性技术	364
2.3 芯片防护技术	364
2.4 物联网技术	365
2.5 环境感知技术	366
2.6 加固防护技术	368
2.7 网电技术	368
2.8 新材料与新工艺技术	369
3 国内研究进展	370
3.1 容错体系结构	370
3.2 软件可靠性技术	371
3.3 芯片防护技术	372

3.4 物联网技术	372
3.5 环境感知技术	374
3.6 加固防护技术	375
3.7 网电技术	375
3.8 新材料与新工艺技术	375
4 国内外研究进展比较	376
5 发展趋势与展望	377
6 结束语	379
参考文献	379
作者简介	383
关键词索引	384
作者索引	386

天地一体化网络研究进展与趋势

CCF 互联网专委会

摘要

天地一体化网络是当前国内外前沿、热点研究领域。本章在分析我国天地一体化网络发展愿景的基础上，给出了天地一体化网络的内涵，进而综述了国内外相关系统的发展情况，并对未来一体化网络的体系结构、组网结构以及路由、传输、星上处理等关键技术的研究难点和发展趋势进行了梳理。

关键词：天地一体化网络，体系结构，组网结构，路由，传输

Abstract

Space and Earth Integrated Network is the frontier and hot research field currently. Based on the analysis of the development vision for our country Space and Earth Integrated Network, the connotation of the integration network has been introduced. Moreover, this paper summarizes the related system development situation at home and abroad. The network architecture, network structure, and the key technology such as routing, transfer of the future integrated network has been introduced and discussed at last.

Keywords: Space and Earth Integrated Network, Network Architecture, Network Structure, Routing, Transfer

1 引言

信息网络是信息获取、传输、处理、分发的基础平台，通过它把各个点、面、体的信息连成一体，从而实现信息的共享和按需使用。

天地一体化网络是综合利用空间和地面等多种技术手段实现信息覆盖的网络系统，目前还没有统一的定义^[1-7]，通过对国内外现状的调研和分析，并结合我国的实际情况，我们认为天地一体化网络的内涵是：

天地一体化网络是以地面网络为基础、以空间网络为延伸，覆盖太空、空中、陆地、海洋等自然空间，为天基、空基、陆基、海基等各类用户的活动提供信息保障的基础设施。

天地一体化网络包括地面网络和空间网络，其中地面网络经过近 20 年的发展之后技术已经比较成熟，同时网络的覆盖也比较全面，但一些偏远地区仍然没有覆盖，通过地面网络进一步完成覆盖的成本过高，同时受国家疆域限制，地面网络无法延伸到国土境外。近年来，随着航天技术的飞速发展，空间网络相关技术日渐成熟，因此加快建设空

间网络，通过天地配合实现网络的延伸，以达到全球无缝覆盖，这一目标将是我国下一步网络发展的重心。

同时，建立天地一体化网络是国家重大战略需求：可满足保卫国家安全的需要，支持现代信息化作战，为多军兵种联合攻防提供信息集成与共享，实现快速反应和精确打击；也可满足保障国计民生的需要，有效整合卫星通信网络、公用通信网络、移动通信网络等，健全应急通信体系，确保突发事件时的信息畅通；还可满足探索空间科学的需要，扩大通信传输覆盖范围，保证信息传输的实时性，为深空探测、移居探索提供支撑。

本章将从我国天地一体化网络的发展愿景出发，结合国内外相关系统的发展现状，分析一体化网络体系结构和组网结构的研究难点、发展趋势，并梳理关键技术的研究现状与发展趋势。

2 我国天地一体化网络的发展愿景

我国建成后的天地一体化网络将具备全球无缝覆盖、一体化互联互通、安全可控等系统能力，为陆、海、空、天等各类用户活动提供全域、全时、安全可靠的信息服务。

2.1 系统能力

天地一体化网络建成后应具备以下能力：

- 覆盖范围：通过天地网络配合，实现全球无缝覆盖；
- 互联能力：支持现有和未来各类网络的接入和互联，包括地面互联网、移动无线网络和空间网络等主流网络；
- 服务对象：支持包括个人、车辆、舰船、飞机、航天器等陆、海、空天各类用户；
- 业务能力：提供以数据业务为主的宽带服务，同时支持语音、图像等特殊业务；
- 安全能力：基于自主知识产权的技术和设备，系统可控可管，提供安全可信的网络信息服务能力；
- 保障能力：天地备份提高网络生存性，提供信息持续保障的能力。

2.2 典型应用

天地一体化网络的典型应用场景有：偏远地区网络接入、空间信息实时传输、海上交通要道信息保障等^[5]，分别介绍如下。

(1) 偏远地区网络接入

虽然我国的地面网络已经比较普及，但受特殊地理条件的限制，一些偏远地区（比如山区、沙漠、海岛等）还没有被地面网络所覆盖，而且通过地面网络进一步完成覆盖的成本过高。建设天地一体化网络，利用天基网络覆盖范围广的优势，则能以相对廉价的方式将这些地区接入到网络中。

(2) 空间信息实时传输

截止 2013 年年底，我国在轨卫星数量已经将近 100 颗，除了通信、导航等系统外，大部分卫星都是对地观测类的，所采集的信息需要传回地面。由于我国没有全球分布的地面站，所以通过地面站回传的方式需要等卫星过顶时才能传输，无法保障信息传输的实时性。只有建设天地一体化网络，通过天基网络实现空间信息传输，才能保障实时传输的要求。

(3) 海上交通要道信息保障

我国至今还没有一个能够向全球提供服务的通信系统，目前海上交通要道的信息保障需要依靠国外的系统，这不仅与我国的大国地位不符，而且存在极大的安全隐患。建设我国自主的天地一体化网络，可以有效地保障我国在海上交通要道的信息服务能力，提升我国的国际地位和综合影响力。

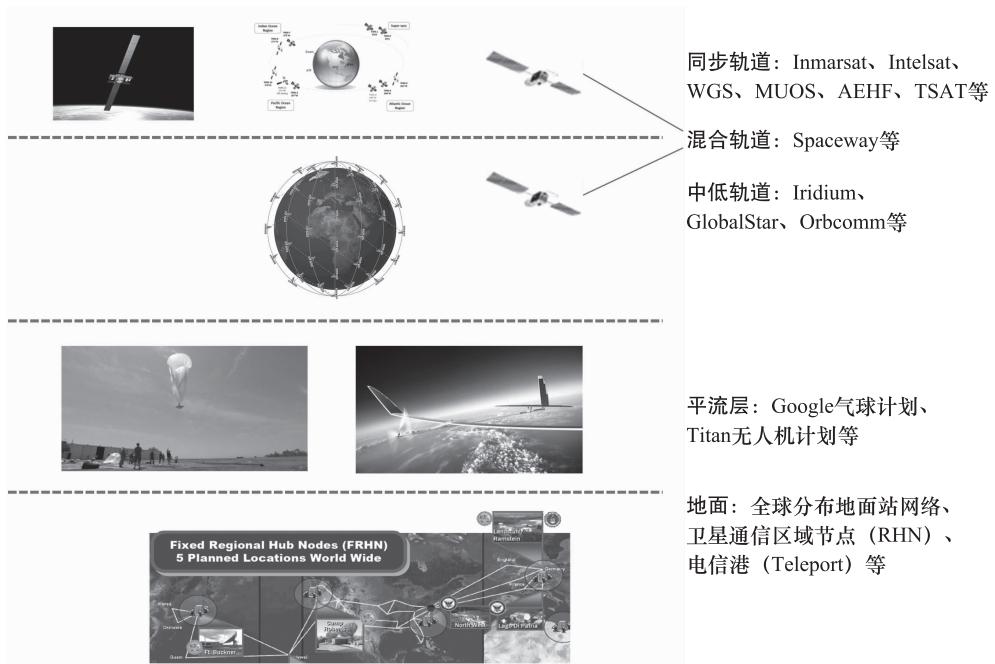
3 国内外现状及差距

3.1 国外发展现状

3.1.1 概述

由于卫星具有高、远、广的独特优势，所以基于卫星的通信方式从一开始就引起了世界各国的广泛关注。但又由于卫星通信涉及的专业和领域多，技术复杂并且建设成本高昂，一个国家卫星通信的技术、发展水平是反映该国综合科技和经济实力的重要标志之一。以美国为例，截止 2013 年年底，美国在轨运行的 440 余颗卫星中，通信卫星有 308 颗（其中静止轨道通信卫星 192 颗）^[8]。

自 1965 年，全球第一颗地球静止轨道通信卫星投入商用，经过半个多世纪的发展，目前全球各国已经建成了众多基于卫星通信的信息系统，实现区域或全球覆盖，如图 1 所示。按照轨道高度分类，基于静止轨道的卫星有：Intelsat、Inmarsat、WGS、AEHF、MUOS 等；基于中低轨道的卫星有：Iridium、Globalstar、Orbcomm 等；另外，还有一类是基于混合轨道的，如 Spaceway（高、中轨道混合）等^[9-11]。



从网络架构来看，可以把不同系统大致归为 3 种模式：天星地网、天基网络、天网地网，如表 1 所示。下面针对这三种组网架构，分别简要介绍国外的几个典型系统。

表 1 国外系统的 3 种组网架构比较

网络架构	典型系统		系统特点
	民用	军用	
天星地网	Inmarsat、Intelsat、Spaceway、Globalstar、Orbcomm 等	WGS、MUOS	星上透明转发，地面组网（全球布站）
天基网络	Iridium	AEHF	卫星空间组网，系统可不依赖地面网络独立运行
天网地网	—	TSAT	空间和地面都要组网，天网、地网两张网、络优势互补

3.1.2 典型系统介绍

1. Inmarsat 系统

国际海事卫星组织 (International Maritime Satellite Service, Inmarsat) 成立于 1979 年，1994 年 12 月更名为国际移动卫星组织，英文缩写保持 Inmarsat 不变。Inmarsat 系统是全球第一个基于地球同步轨道卫星的全球卫星移动通信系统，也是目前全球最大的卫星移动通信系统。经过 30 多年稳定良性发展，Inmarsat 现已发展成为覆盖地球约 85% 土地（除南北两极）和大约 98% 人口的卫星移动通信系统，向全球的海上、陆地、航空用户提供话音、数据、互联网接入以及海上搜救业务。截至 2013 年年底^[11]，Inmarsat 系统

在轨运行的卫星有 11 颗（包括 1 颗 Inmarsat-2、5 颗 Inmarsat-3、4 颗 Inmarsat-4、1 颗 Inmarsat-5）。

（1）Inmarsat 系统发展历程

Inmarsat 系统发展至今，已经经历了五代，目前在使用的是第四代，并且正在开始建设第五代，整个发展历程简要介绍如下。

1) Inmarsat-1 系统。Inmarsat 成立时，没有属于自己的卫星。只能租用美国通信卫星公司（COMSAT）的 Marisat 卫星、欧洲宇航局的 Marecs 和国际通信卫星组织的 Intelsat-V 卫星，运营 Inmarsat-1（第一代 Inmarsat）系统。

2) Inmarsat-2 系统。Inmarsat 于 1991 年 3 月 8 日、10 月 30 日、11 月 29 日和 1992 年 4 月 5 日，发射了四颗卫星，构成了 Inmarsat-2 卫星星群，分别为大西洋东星，定位于 15.5°W ；印度洋星，定位于 64.5°E ；太平洋星，定位于 178°E ；大西洋西星，定位于 54°W 。Inmarsat-2 卫星为全球波束卫星。

3) Inmarsat-3 系统。1996 年 4 月，所发射的第 1 颗 Inmarsat-3 卫星被部署于印度洋上空，1996 年 9 月发射的第 2 颗 Inmarsat-3 卫星覆盖大西洋东，1996 年 12 月 17 日发射的第 3 颗 Inmarsat 卫星覆盖大西洋西，1997 年 6 月 3 日发射的第 4 颗卫星覆盖太平洋，它们构成 Inmarsat-3 卫星星群，Inmarsat-3 是 Inmarsat 第一次采用区域波束，并将区域波束和全球波束结合在一起使用的系统。Inmarsat-2、Inmarsat-3 都使用 4 颗卫星覆盖全球。

4) Inmarsat-4 系统。2005 年 3 月 11 日，第 1 颗 Inmarsat-4 代卫星发射，起初将其定位于 64°E ，覆盖印度洋地区。2005 年 11 月 8 日，所发射的第 2 颗 Inmarsat-4 卫星定位于 53°W ，覆盖美洲地区。2008 年 8 月 18 日，所发射的第 3 颗 Inmarsat 卫星定位于 98°W ，覆盖美洲。2009 年 1~2 月间，有两颗卫星完成变轨，分别定位于 25°E , 143.5°E 。3 颗 Inmarsat-4 卫星覆盖全球。Inmarsat-4 卫星在功率、容量、适应性 3 方面开创了移动卫星通信的新纪元。与上一代 Inmarsat-3 卫星相比，每颗 Inmarsat-4 卫星的功能强大了 60 倍，容量也大了 20 倍。与前几代卫星不同，Inmarsat-4 首次使用 3 颗卫星覆盖全球。

5) Inmarsat-5 系统。2008 年，在 Inmarsat-4 投入运营还不到 3 年时，Inmarsat 就着手研制下一代海事卫星系统 Inmarsat-5，以构建全球宽带无线网络（又称 Global Xpress）。Inmarsat-5 卫星采用波音公司成熟的 702HP 卫星平台，星上有 89 个 Ka 转发器，额定功率为 15kW ，设计使用寿命为 15 年。2013 年 12 月首颗 Inmarsat-5 卫星发射成功，2014 年年底完成了 3 颗卫星的发射并形成全球覆盖，2015 年开始提供服务。

Inmarsat-5 将是第一个使用 Ka 波段的商用全球网络系统，采用 Ka 波段点波束，可使终端卫星天线口径小至 20cm（差不多一个 iPad 大小）。根据 Inmarsat 官网介绍，Inmarsat-5 可以为 60cm 移动终端提供最高为 50Mbps 的下行速率和最高为 5Mbps 的上行速率，整个 Inmarsat-5 系统的吞吐率将达 100Gbps 。

各代 Inmarsat 系统的技术参数情况如表 2 所示，随着技术的进步和商业模式的拓展，卫星的服务能力不断提升，用户速率也不断提高，提供的业务从语音、中低速数据逐渐扩展到宽带服务。

表2 各代 Inmarsat 系统的技术参数比较

技术参数	Inmarsat-2	Inmarsat-3	Inmarsat-4	Inmarsat-5
启动时间	1990年	1996年	2005年	2013年
全球服务时间	1992年	1998年	2009年	2015年
卫星数量	4	4+1	3+1	3*
波束	1个全球波束	1个全球波束；7个区域波束	1个全球波束；19个区域波束；193个点波束	—
卫星发射重量/kg	1500	2050	5959	—
用户速率	4.8kbps	64kbps	394kbps	5Mbps
业务	话音、低速数据	话音、中速数据	基于IP的宽带数据业务（陆地BGAN、海上FB、航空SB）	全球宽带（Global Xpress）

注：首颗 Inmarsat-5 卫星已于 2013 年 12 月发射成功。

(2) Inmarsat 系统介绍

目前正在运行的是 Inmarsat-4 系统，采用天星地网的组网架构，其中天上部分由 3 颗同步轨道卫星实现全球覆盖，另有 1 颗备用卫星，3 颗卫星的在轨位置分别是：亚太卫星为东经 143.5 度 (143.5°E)，覆盖亚洲和西太平洋区域；欧非卫星为东经 25 度 (25°E)，覆盖欧洲、中东和非洲区域；美洲卫星为西经 98 度 (98°W)，覆盖美洲、大西洋和东太平洋区域。3 颗卫星基本覆盖了全球 $78^{\circ}\text{S} \sim 78^{\circ}\text{N}$ 之间的区域，3 颗卫星的全球覆盖情况如图 2 所示。

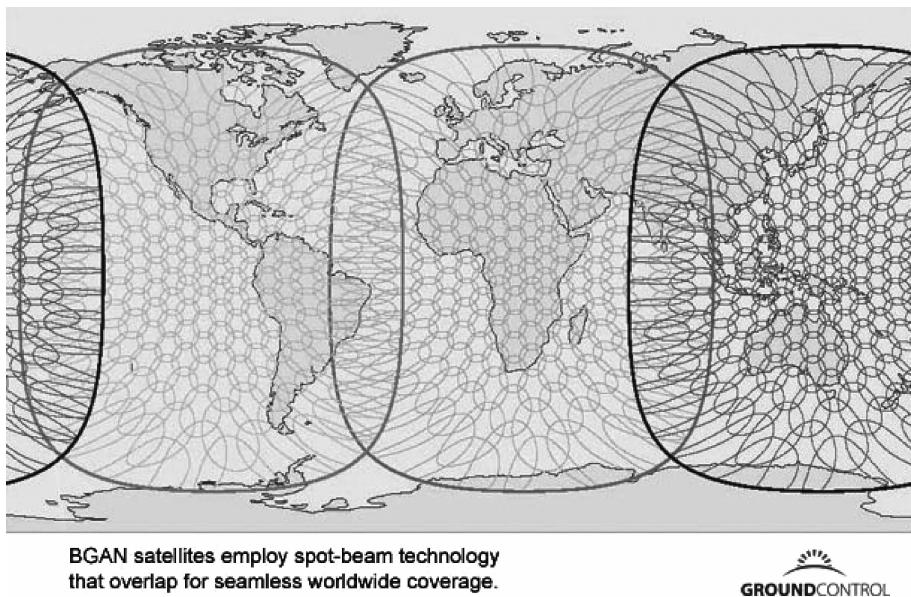


图2 Inmarsat-4 系统覆盖图

Inmarsat-4 系统的地面对部分主要由全球分布的卫星关口站 (SAS)、网络汇接中心 (MMP) 和地面接续站 (POP) 等组成，如图 3 所示。

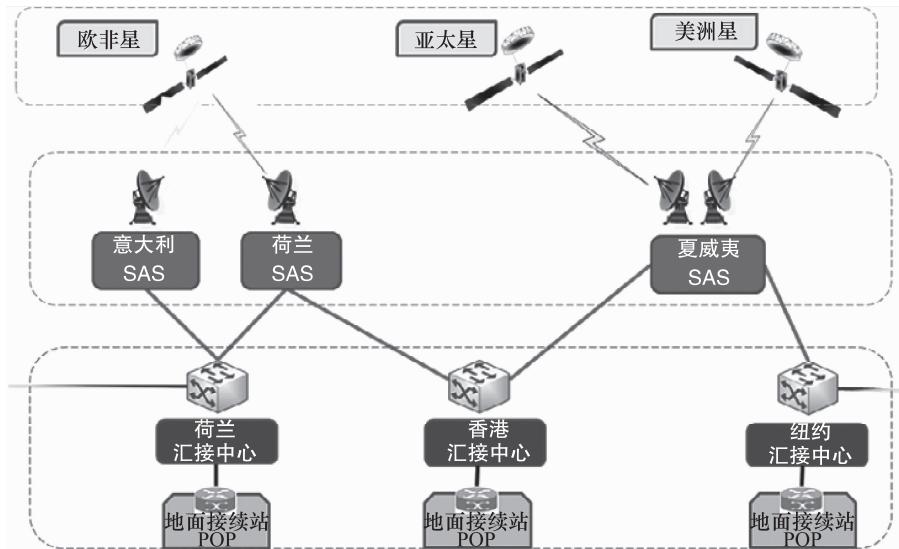


图 3 Inmarsat 系统网络架构示意图

卫星关口站（SAS）是卫星和陆地网络通信的关节点，负责处理用户终端的业务申请、交换，分配用户资源、容量等，提供语音、数据、视频等通信。Inmarsat-4 系统对关口站的布局采取了新的架构，目前全球只设有 3 个关口站，位于荷兰布卢姆和美国夏威夷的为主用关口站，位于意大利佛希罗的为备用关口站。中国区域的通信都要经过美国夏威夷关口站转接。2011 年 12 月 8 日，交通运输部中国交通通信信息中心第四代国际海事卫星北京关口站工程（简称“北京关口站工程”）启动，其建成后将成为全球第三个投入运营的主用关口站和 Inmarsat-4 业务全球网络的重要组成部分。

为了使 Inmarsat-4 业务能够满足陆地用户的接入要求，系统分别在美国的纽约、荷兰的阿姆斯特丹和中国的香港 3 个通信网络发达的城市建设了网络汇接中心，各地业务提供者通过数据网络地面接续站提供具体的专线接入服务和专业化应用。目前北京 Inmarsat-4 地面接续试验网与中国香港的 MMP 相连，并且我国还分别在北京和中国香港建设了 POP 站。

第四代海事卫星改变了前三代海事卫星系统完全使用传统电路交换技术的现状，增加了数据分组交换技术，并在其关口站引入了公共移动通信网络的 3G 技术，采用 3GPP 标准，实现高达 492 kbps 的数据带宽，支持 3G 语音、数据和视频传输等功能。

第四代海事卫星宽带业务主要实现以下基本功能：电路语音（AMBE2+）、ISDN（64 kbps 数据和 3.1 kHz 音频）、标准 IP 业务（492 kbps）、6 个等级的流媒体 IP 业务（最高等级为 384 + kbps）。这些基本业务可支持的应用主要有：国际国内双向电话、传真、短信、语音邮箱、连接互联网的数据传输、连接专用网的数据传输和视频传输等。

陆地宽带业务（Broadband Global Area Network，BGAN）终端：BGAN 终端分为便携式和车载式两种类型。BGAN 已被纳入国家应急平台体系，可在防灾减灾、突发公共事件、特殊任务、国际救援、局部战争、恐怖袭击等各种情况下提供语音、数据和视频应用。

海上宽带业务（Fleet Broadband，FB）终端：安装在船舶上的卫星设备，其天线能自动跟踪卫星。FB 是全球海上遇险与安全系统（GMDSS）的组成部分，支持海上安全信息传输（航行及气象预警、位置等）、搜救协调通信和海上反恐等应用。

航空宽带业务（Swift Broadband，SB）终端：安装在飞机上的机载卫星通信终端，目前正在履行满足航空通信安全的程序，提供飞机上语音、数据和视频的多种应急通信。

（3）Inmarsat 系统对我们的启示

从 Inmarsat 系统的发展历程来看，可以总结出以下几个特点。

1) 同步轨道。同步轨道卫星的优势是覆盖范围广，用几颗卫星就能实现准全球覆盖（两极地区除外），第二、三代 Inmarsat 采用 4 颗卫星实现全球覆盖，第四代之后开始采用 3 颗卫星实现全球覆盖。卫星数量的减少所带来的明显好处是系统建设成本低，管理维护也比较简单。

2) 天星地网。采用天星地网的网络架构的好处是天上的卫星之间不用组网，星上不需要复杂的处理，因此可以把有限的资源用于扩大系统容量。但是另一方面，这种网络架构对地面网络提出了更高的要求，需要在全球范围内布设地面关口站，以实现整个系统的网络互连。因为我国不具备全球分布的地面关口站网络，所以不建议采用这种网络架构来构建我国的天地一体化网络。

3) 迭代发展。采用迭代发展的思路是，从最简单的可用系统起步，随着技术的进步和商业模式的演进，逐步升级系统功能和容量、扩展业务范围，实现商业上的良性持续发展。Inmarsat 系统一开始只是为海上用户提供话音和低速数据服务，经过几代发展之后，扩展到了陆地和空中用户，同时业务能力也向宽带化发展。

2. Iridium 系统

Iridium（铱星）卫星通信系统（以下简称“铱星系统”或“铱星”）使人类实现了在地球上任何地方都可以相互联络的“神话”，即实现了 5 个任何（5W）：任何人（Whoever）可以在任何地点（Wherever）任何时间（Whenever）与任何人（Whomever）以任何方式（Whatever）进行通信。铱星是世界上第一个投入使用的大型低轨道移动的通信卫星系统，它开创了全球个人通信的新时代，被公认为是现代通信史上的一个里程碑^[4]。

（1）Iridium 系统发展历程

Iridium 系统（见图 4）是美国摩托罗拉公司（Motorola）于 1987 年设计的低轨道全球个人卫星移动通信系统，它与现有通信网结合，可实现全球数字化个人通信。该系统历经 11 年筹建、耗费 50 亿美元，于 1998 年 11 月正式投入运营。

尽管技术具有先进性，并且系统也顺利建成和启用，但由于在此期间地面蜂窝移动通信系统的飞速发展，Iridium 系统一直无法正常发展壮大，再加上高昂的建设和维护费用，该系统不能实现扭亏为盈，被迫于 1999 年 8 月 13 日进入破产保护。Iridium 系统在此后的一段时间内一直乏人问津，直到 2000 年 12 月，若干投资者组成的 Iridium Satellite LLC 公司全盘接收了 Iridium 系统，并随后获得了美国国防部授予的合同订单，Iridium 系统才走上了重生之路。2001 年 3 月，Iridium 系统重新恢复了通信业务，此后逐渐发展壮大，到 2009 年 6 月，Iridium 卫星手机用户已达到 34.7 万人。

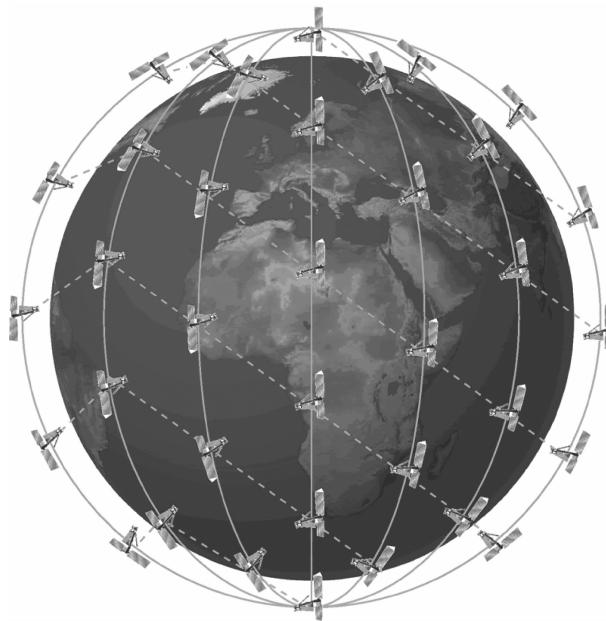


图 4 Iridium 系统卫星星座示意图

2007 年 2 月，铱星公司宣布了 NEXT 计划，在保持原有星座架构的基础上提供更大容量和更高数据率的业务，L 频段业务高达 1.5Mbps，同时支持高达 8Mbps 的高速 Ka 频段业务。

2009 年，美国空军投资 Iridium NEXT 研究，分析 NEXT 星座如何搭载空间目标监视传感器，为全球范围内的美军作战平台提供实时通信及空间目标监视的综合能力。从目前的发展来看，Iridium NEXT 已经确定搭载具备气候变化监视、多光谱对地成像、空间气象监视能力的载荷。

2015 年，Iridium NEXT 星座开始部署，美军全球范围内的作战单元都将具备实时通信、空间目标监视以及导航定位等综合能力。

(2) Iridium 系统介绍

Iridium 系统星座由 66 颗围绕 6 个极地圆轨道运行的 LEO 卫星组成，轨道高度约 780km，每颗卫星重 386kg。卫星有星上处理器和星上交换功能，并且采用了星间链路，因而系统的性能极为先进。

每颗卫星拥有 4 条 Ka 频段（22.55 ~ 23.55GHz）的星间通信链路。Iridium 星体呈三面棱体结构，3 个 Ka 频段的阵列天线板安装在 3 条棱线之上，但 Iridium 可以同时产生同轨面前后两个、异轨面左右两个星间链路波束，其 4 个星间链路可由 3 个 Ka 频段的阵列天线根据需要切换产生。两条用于建立同轨面前后方向卫星的星间链路，两条用于建立相邻轨面间卫星的星间链路，星间距离为 2700 ~ 4400km，数据速率可达 25Mbit/s。异轨面间链路的天线可根据加载到卫星上的星历信息进行指向调整，波束宽度足以适用纬度控制和卫星位置保持的容差。

Iridium 系统的最大特点是，通过卫星之间的接力来实现全球通信，相当于把地面蜂

将移动电话系统搬到了天上。它与目前使用的静止轨道卫星通信系统比较有两大优势：一是轨道低，传输速度快，信息损耗小，通信质量高；二是 Iridium 系统不需要专门的地面接收站，每部移动电话都可以与卫星联络，这就使地球上人迹罕至的不毛之地、通信落后的边远地区以及自然灾害现场都变得畅通无阻。因此，Iridium 系统开启了个人卫星通信的新时代。

（3）Iridium 系统对我们的启示

从 Iridium 系统的发展历程来看，可以总结出以下几个特点。

1) 低轨星座。利用低轨卫星星座的主要优势有两个：一是实现真正意义的全球覆盖（包括南北极地区）；二是距离地面近、时延小、终端可以做到小型化。然而，低轨卫星星座需要的卫星数量多，至少需要数十颗卫星，同时卫星相对地面高速移动，系统建设和维护的复杂度高。

2) 空间组网。Iridium 系统的每颗卫星都拥有 4 条星间链路，通过星间链路可实现空间组网，构成完整的天基网络。这样做的好处是不需要全球分布的地面站系统，但由此产生的问题是，整个系统的技术复杂度高，建设和运营的成本较高，容易导致商业上的失败。

3) 多功能综合。由于 Iridium 系统建设和运营的成本较高，单一的通信业务与地面网络相比不具备成本优势，下一代系统（Iridium NEXT）逐渐向多功能综合发展，在卫星上搭载多种载荷（包括远程传感、气象监测、对地观测等）。

3. WGS 系统

WGS 系统是美军的宽带卫星通信，其前身是 DSCS 系统，主要用于解决大容量、高宽带、高数据率的干线通信需求，主要频段为 X 频段和 Ka 频段^[13]。

（1）WGS 系统发展历程

WGS 系统是美国新一代静止轨道军用通信卫星系统，用于替代国防卫星通信系统（DSCS），为美军提供宽带卫星通信业务，卫星主承包商是波音公司。该系统分三个阶段部署：第一阶段包括 3 颗卫星，分别于 2007 年 10 月 10 日、2009 年 4 月 4 日和 2009 年 12 月 6 日发射；第二阶段包括 3 颗卫星，分别于 2012 年 1 月 20 日、2013 年 5 月 24 日和 2013 年 8 月 7 日发射；第三阶段包含国际合作，卫星具体数量尚未确定，截至目前，美军已经签署了 WGS-7 到 WGS-10 的合同，同时正在开展 WGS-11、WGS-12 的合同谈判。

（2）WGS 系统介绍

WGS 系统如图 5 所示，该系统比较显著的特点是：①大容量，瞬时可交换带宽为 4.875GHz；②星载数字信道化技术，支持不同波束、不同频段（X、Ka 频段）子信道间的星上交换；③支持高速 AISR 应用。

第一阶段 Block-I 系列 3 颗卫星已经入轨，覆盖了太平洋、中东、欧洲以及非洲。每颗卫星的容量可达 2.4 ~ 3.6Gbps，与原有系统相比，通信容量提高 10 倍以上。

第二阶段 Block-II 系列除了具备 Block-I 卫星的特点外，还增加了一种射频旁路能力，可以将“全球鹰”ISR 数据传输能力从 137Mbps 提高到 274Mbps。2012 年 1 月，Block-II 系列的首颗 WGS-4 升空，该星将用以保障美军部署在近东和亚洲地区的无人机通信，其中包括美军广域海上监视（BAMS）无人机和“全球鹰”无人机。

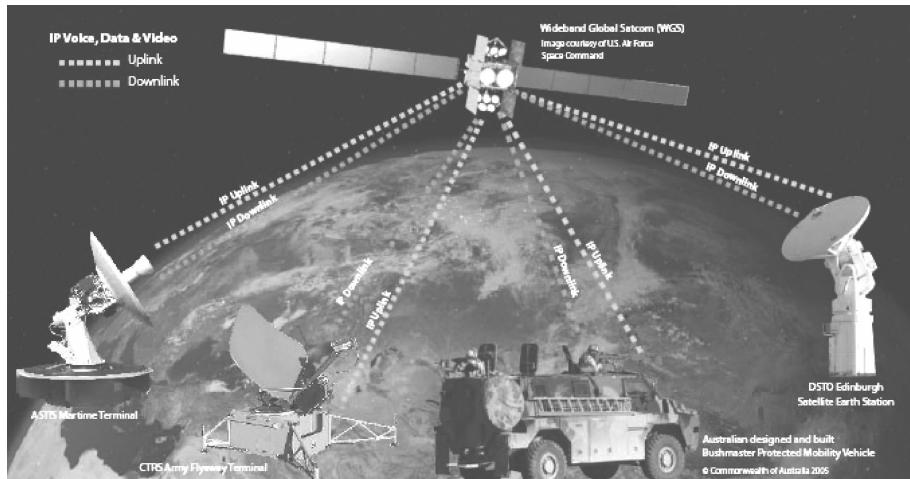


图 5 WGS 系统示意图

(3) WGS 系统对我们的启示

从 WGS 系统的发展历程来看，可以总结出以下几个特点。

- 1) 同步轨道。采用同步轨道的优点与前面介绍的 Inmarsat 系统类似，即只用几颗卫星就可以实现全球覆盖。
- 2) 天星地网。WGS 系统主要服务于宽带用户，依托于美军在全球分布的卫星通信区域节点（RHN）及电信港（Teleport）等设施，从而实现系统互联，因此不需要星上组网，对于采用天星地网的网络架构，我国受国情限制没法照搬美国的这种网络架构。
- 3) 迭代发展。基于前期 DSCS 系统的建设运营经验和技术积累，WGS 系统逐步发展，其现有传输速率、容量相比 DSCS 系统大幅提升，一颗 WGS 卫星的容量比整个 DSCS 系统的总容量还要大。

4. AEHF 系统

AEHF 卫星系统是美国新一代高防护性能的地球静止轨道军事通信卫星系统，用于替代老化的“军事星”（Milstar）卫星系统，在包括核战争在内的各种规模战争中，为美军关键战略/战术部队提供防截获、抗干扰、高保密和高生存能力的全球卫星通信^[9]。

(1) AEHF 系统发展经验

AEHF 系统的前身是美军的 Milstar 系统，该系统是美海、陆、空各军种的一项联合任务计划，是美国为了确保冷战时期核战争条件下的三军保密通信，于 20 世纪 80 年代初开始实施的一项军事卫星通信系统工程。它原是美苏对峙冷战时代的产物，在所谓的“星球大战”计划中，曾被视为在美国战略系统中绝对位居首位的计划，也是美国政府实现战略/战术部队现代化的一个关键要素。

Milstar 系统以方便的呼叫方式为部队，尤其是为大量战术用户提供实时、保密、抗干扰的通信服务，通信波束覆盖全球。Milstar 系统采用了 EHF 频段、跳频（速率达每秒几千次）、自适应天线调零处理、灵活可变的波束覆盖、完善的星间（ISL）链路、较强的星上数字信号处理等技术，实现了低截获概率（LPI）和低检测概率（LPD），达到了

很好的通信抗干扰效果。

1994年3月，首颗 Milstar (DFS-1) 卫星发射入轨，定点在 120°W ，1995年11月，第二颗 Milstar (DFS-2) 卫星顺利升空，后定点在 4°E 。DFS-1、DFS-2 两星配对工作，提供对美太平洋至大西洋部队的保密通信覆盖，并获得了成功。Milstar-1 卫星的设计寿命是 7 年以上，质量为 4500kg，主要携带低速数据率通信载荷 (LDR)，可发送、接收速率为 $75 \sim 2\,400\text{bps}$ 的声码和数据信息 (无图像)。

第一代 Milstar 卫星投入应用的成功激发了美军发展第二代 Milstar 卫星的积极性，4 颗 Milstar-2 (DFS-3、4、5、6) 卫星在 2002 年之前全部发射升空，从而形成全球覆盖的抗干扰卫星通信网。与 Milstar-1 不同，Milstar-2 卫星在轨寿命达 10 年以上，携带中速数据率有效载荷 (MDR)，目的是为战术部队提供大容量的数据 (如图像) 传输能力，MDR 的最高传输速率为 50Mbps ，每个用户的接入速率为 1.544Mbps 。

AEHF 系统是 Milstar 系统的后继型号，于 20 世纪 90 年代中期提出，1999 年 4 月正式启动。它的总目标是在静止轨道上放置 AEHF 卫星，为现有的 Milstar 和将来使用 EHF 的用户提供全世界范围的安全保密、生存力强、后台兼容、易于操作、便于监控的通信方式。覆盖南纬 65° 到北纬 65° 的地球表面，比第一代 Milstar 数据速率快 10 倍。计划发射 6 颗卫星，目前已发射 3 颗 (分别于 2010 年 8 月、2012 年 5 月、2013 年 9 月发射)。

(2) AEHF 系统介绍

AEHF 系统由天上卫星、地面任务控制中心和用户终端组成，如图 6 所示，天上的多颗卫星之间采用毫米波星间链路，实现空间组网，可为处在各种军事行动冲突级别的作战人员提供联合的、可互操作的和有保证的连接能力。

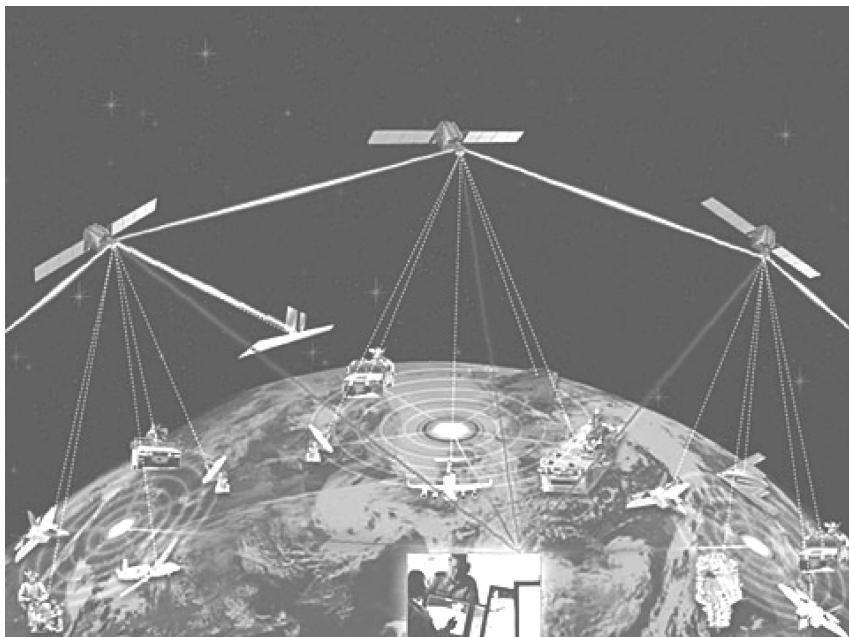


图 6 AEHF 系统示意图

AEHF 卫星采用洛马公司的 A-2100M 平台，设计寿命为 15 年，发射质量约 6 600kg，入轨质量为 4 100kg，比采用波音公司 BSS-702 平台的“宽带全球卫星通信”卫星大 10% 以上。

AEHF 卫星在 Milstar-2 卫星低数据率载荷和中数据率载荷的基础上，增加了扩展数据率载荷（XDR），最大数据传输速率提高到 8.192Mbps。AEHF 卫星一共携带有 14 副天线，单星通信总容量从 Milstar-2 的 40Mbps 提高到 430Mbps。与 Milstar-2 卫星相比，AEHF 系统的星间链路增强了路由功能和抗干扰能力，卫星总容量提高了 10 倍，星间链路的数据传输速率也由 10Mbit/s 提到了 60Mbit/s。AEHF 卫星能够同时支持 2 000 部用户终端，可创建的通信网络数量由 Milstar-2 的 1 500 个增加到 4 000 多个。

AEHF 卫星采用了星间链路、星上处理等先进技术，能根据用户优先级别来提供点对点通信及网络通信服务。该系统有非常强的战场生存能力，即便地面控制站被破坏，整个系统仍能自主工作半年以上。

（3）AEHF 系统对我们的启示

从 AEHF 系统的发展历程来看，可以总结出以下几个特点。

1) 同步轨道。采用同步轨道的优点与前面介绍的 Inmarsat 系统类似，只用几颗卫星就可以实现全球覆盖。

2) 空间组网。AEHF 系统主要用于满足军事上的抗干扰和高生存性要求，采用星间组网、星上处理和卫星自主控制技术，具有很强的生存能力，可以在不依赖地面支持的情况下自主运行半年以上。

3) 迭代发展。基于前期 Milstar-1/2 两代系统的建设运营经验和技术积累，AEHF 系统逐步发展，其现有的传输速率和容量相比前两代大幅提升，但仍保持对之前系统的兼容。

5. TSAT 计划

为了跟上“新军事变革”的发展，适应“网络中心战”的要求，美军于 21 世纪初提出了“转型通信体系结构”（TCA）。TCA 的目的是提供一套受保护的、类似互联网的安全通信系统，把太空、空中、陆地、海洋的网络整合为一体化，从根本上改善美军的通信能力。这一结构由转型卫星通信系统（TSAT）、全球信息栅格宽带扩展计划（GIG-BE）、战术级作战人员信息网（MN-T）和联合战术无线电系统（JTRS）等 4 大计划组成。其中，TSAT 计划是 TCA 体系结构的核心，发挥重要的组网和通信作用，提供全球性、高带宽、高安全性、自动化及动态能力^[9]。

（1）TSAT 计划发展历程

2004 年，TSAT 计划开始实施，并计划于 2011 年发射第一颗卫星。2008 年，国防部批准了 TSAT 的下一阶段计划，并计划建造 5 颗卫星和 1 颗备份星。但由于关键技术成熟度和经费预算超支等，TSAT 计划于 2009 年被取消。尽管如此，TSAT 计划的很多先进理念还是被继承了下来，并体现在后续项目中。

带宽是导致 TSAT 计划产生的直接原因。为填补 TSAT 计划取消后导致的抗毁、抗干扰的可靠通信能力空白，美军考虑增加 WGS 及 AEHF 卫星的部署数量，其中 WGS 卫星

达到 12 颗，AEHF 卫星达到 6 颗，以此来弥补不断上升的通信容量需求，保护通信能力。

(2) TSAT 计划介绍

TSAT 计划由美国空军主持，旨在建立一个服务于美国国防部（即美国军方）、美国航空航天局（NASA）和美国情报界（IC）的保密、高容量的全球卫星通信网络。除了极大提高信息容量和安全性外，TSAT 还会采用一些新技术，如卫星光通信技术、IPv6、星载路由技术、大孔径天线等。这些技术能发挥以下作用：一是可以实现军事用户使用小型终端进行“动中通”（COTM）和军事用户全球性动态链接；二是 TSAT 可以从受保护的空基和天基情报、侦察、监视（ISR）资源中获取信息，增强系统态势感知能力；三是 TSAT 能提供高容量的综合信息分享能力，以前所未有的规模实现网络环境下的互操作。

TSAT 计划主要用于在太空建立类似地面的 Internet 网络，为地面用户、空中及太空武器平台的信息传输提供太空路由，保证数据能够以最直接的方式从信源传送到信宿，增强端用户获取信息的时效性，从而提高作战部队抢先发现目标、快速决策和及时评估战果的能力。如图 7 所示，TSAT 系统的天上部分是由 5 颗卫星组成的星座，提供 EHF、X、Ka 频段，直接和非直接地与 AEHF、MUOS、WGS 和 APS 连接，地面部分接入全球信息栅格（GIG）网络。

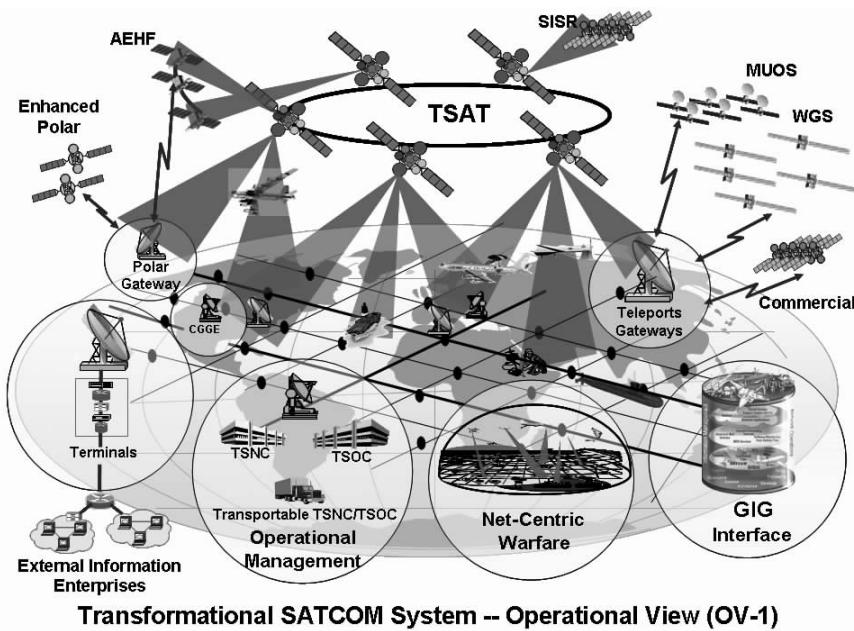


图 7 TSAT 系统示意图

(3) TSAT 计划发展经验

从 TSAT 计划的发展历程来看，可以总结出以下几个特点。

1) 同步轨道。TSAT 计划采用 5 颗同步轨道卫星形成天基骨干网，同步轨道的优势如前所述，此处不再重复。

2) 天网地网。TSAT 计划采用天网地网的架构，其中天基网络由 5 颗同步轨道卫星通过星间链路构建而成，地面接入美军的全球信息栅格（GIG），依托美军在全球设立的卫星通信区域交换节点（RHN）和电信港等，实现全球互连。

3) 网间互连。TSAT 是美军构建以网络为中心，实现空间系统、地面网络、终端系统互连互通，建立兼容、可互操作的“系统的系统”的一次尝试。TSAT 计划打造可以连接现有多个系统的空间骨干网络，虽然最终该计划因为巨大经费等被推迟，但其先进理念仍在后续项目中得到继承，比如太空路由器（IRIS）等。

3.1.3 总结

在此对国外典型系统的发展情况进行了汇总，如表 3 所示。

表 3 国外典型系统总结

系统名称	轨道类型	网络架构	系统特点	对我们的启示
Inmarsat	同步轨道	天星地网	迭代发展：用户从海上扩展到陆地、空中，业务从话音扩展到宽带	天星地网的架构不适合我国，迭代发展的思路值得借鉴
Iridium	低轨星座	天基网络	星间组网简化地面系统，并从单一通信功能向通信、导航、对地观测等多功能综合发展	低轨星座的技术较复杂，建设和维护成本高，不建议发展单一功能的低轨通信系统
WGS	同步轨道	天星地网	大容量、高宽带、高数据率的干线通信	天星地网的架构不适合我国
AEHF	同步轨道	天基网络	星间组网，可不依赖地面网络而自主运行，抗干扰、军事顽存	星间组网技术值得借鉴，但不必为了强生存性使星上设计太复杂
TSAT	同步轨道	天网地网	空间和地面都组网，以网络为中心互连多个系统	理念值得借鉴，技术上可逐步迭代升级

综上所述，可以得出以下几点结论：

- 1) 比较成功的系统大多数都是经过迭代发展的，先从简单系统开始，随着技术进步和运营经验的积累逐步升级、完善；
- 2) 从技术成熟度和商业化运营角度看，基于同步静止轨道卫星的系统相比基于中低轨卫星星座具有优势；
- 3) 星间链路技术趋于成熟，容量不断提升：其中 Ka 波段链路已经成熟，激光链路也正在试验中；
- 4) 构建可以实现不同系统互连互通的天地一体化信息基础设施是未来的发展趋势；
- 5) 从网络架构来说：天星地网架构的技术比较成熟，应用广泛，但不适合我国；天基网络架构在安全性、抗毁性和独立性方面有优势，但因为要考虑脱离地面网络独立运行，加大了对星上处理和星间信息传输能力的要求，导致技术复杂，系统的建设和维护成本高，在商业上难以成功；天网地网架构通过天地两张网络的配合，充分利用天基网络的广域覆盖能力和地面网络丰富的传输、处理能力，降低了整个系统的技术复杂度和成本。

3.2 国内发展现状

3.2.1 概述

我国人造地球卫星主要包括科学探测与技术试验卫星、气象卫星、对地观测卫星、通信广播卫星、中继卫星、定位卫星等，且已由试验阶段进入了应用阶段。截至 2013 年年底，我国在轨卫星数量已接近 100 颗（全球在轨卫星数量突破 1 000 颗，美国拥有近 500 颗）。

同时，我国地面网络蓬勃发展，在用户规模、信息资源、产业规模等方面均处于世界前列，在国民经济和社会发展中发挥着越来越重要的作用。

伴随着我国社会、经济快速发展，特别是各种新业务、新应用的出现，对空间信息网络提出了众多新需求。如何更好地发展我国空间和地面的资源成为了一个十分值得研究、探索的重要议题。

3.2.2 地面网络发展现状

近年来，我国包括互联网、移动通信网在内的地面网络得到了迅猛发展。

(1) 网络规模位于世界前列，国际互连能力不断增强，已能满足绝大部分应用需求^[14]

在网络用户数方面，互联网网民的规模实现跨越增长，2008 年我国网民规模已跃升全球第一，2013 年 12 月达到 6.18 亿，继续位居世界第一位，互联网普及率为 45.8%，超过世界平均水平。

在 IP 地址数目方面，截至 2013 年 12 月，我国 IPv4 地址总数基本维持不变，共计有 3.30 亿个，位列世界第二位；我国 IPv6 地址数量为 16 670 块/32，较 2012 年同期增长 33.0%，位列世界第二位。

在基础设施建设方面，我国已建成辐射全国的通信光缆网络，总长度达 826.7 万公里，其中长途光缆线路达 84 万公里。我国 99.3% 的乡镇和 91.5% 的行政村接通了互联网，96.0% 的乡镇接通了宽带。2009 年 1 月，我国开始发放 3G 牌照，截止 2013 年 11 月底，全国移动电话用户达 12.23 亿户，位居世界第一位，其中 3G 用户有 3.87 亿户，占比达 31.6%。

在国际互联方面，截至 2013 年 12 月，我国国际出口带宽已超过 3 400 Gbps，较 2012 年增长了 79.3%（见图 8）。

中国互联网的发展已从“普及率提升”进展为“使用质量提升”：首先，国家政策支持，2013 年国务院发布《国务院关于促进信息消费扩大内需的若干意见》，说明了互联网在整体经济社会的地位；其次，互联网与传统经济结合得越加紧密，如购物、物流、支付乃至金融等方面均有良好应用；再者，互联网应用逐步改变人们的生活形态，对人们日常生活中的衣食住行均有较大改变。

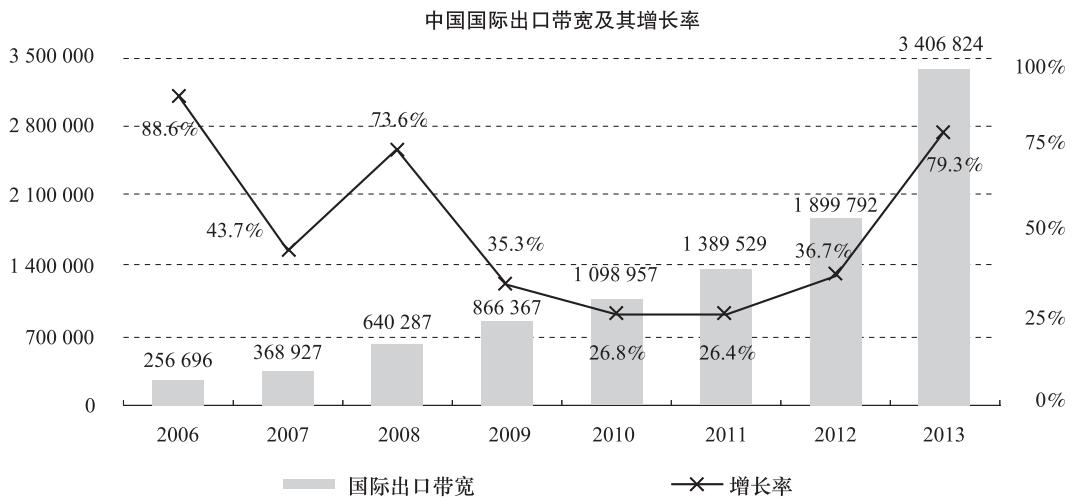


图 8 我国国际出口带宽增长情况

(2) 网络技术先进，部分下一代互联网技术已达世界领先水平

我国较早开展了下一代互联网的研究，实施了一系列国家级技术创新计划、应用示范和试商用工程，并已经取得了举世瞩目的成绩。在下一代互联网建设方面，通过实施中国下一代互联网示范工程 CNGI^[15]，建成了大规模的下一代互联网示范网络，这一网络包括 6 个主干网、2 个国际交换中心、273 个驻地网，其正在发挥技术创新、业务创新、试验验证、应用示范、规模使用等方面的重要作用。其中，清华大学等 25 所高校建成的 CNGI-CERNET2 是目前世界上规模最大的纯 IPv6 互联网，涉及多项重大创新，总体上达到世界领先水平。

在设备研发方面，我国基本掌握了 IPv6 关键网络设备的核心技术，路由器、交换机、宽带接入设备、互联网关、音/视频监控摄像终端、无线传感器网络节点等产品已经批量投入市场，设备研发和产业化能力达到国际先进水平。

在业务应用方面，我国已经开展业务系统的研发以及新业务的应用示范，且部分已经在我国经济和社会建设中发挥了积极作用。例如北京奥运会期间，CNGI 北京互联中心开通 IPv6 奥运官网镜像站点，这是我国面向全球的 IPv6 的重要应用示范，期间，IPv6 视频监控和传感器系统成功应用于全国各地的 48 个体育场馆，圆满完成奥运的通信保障任务，在国际上引起了很大反响。

在技术创新方面，我国已经在网络过渡、安全机制等方面取得了局部突破，在国际互联网标准化工作方面取得了长足进展，在国际互联网标准化组织 IETF 中的主动权和话语权日益扩大。目前我国专家主导制定了 40 多项 IETF 标准，这从一定程度上改变了我国长期以来在互联网核心技术方面受制于人的被动局面。不过值得指出的是，我国在对核心技术和核心标准的掌控力方面与发达国家相比，特别是与美国相比，差距还很大。

我国地面网络的发展同时也存在一定的问题，主要体现为如下几点。

(1) 境内难以实现“全”覆盖，对境外网络无控制能力

地面网络在建设过程中受到国土范围、物理环境等因素的限制，存在着境内无法全面覆盖的问题。我国有着辽阔的海洋面积，内海和边海的水域面积约 470 多万平方公里，然而建设地面基础设施的方式无法用于实现对海洋的覆盖。另外，对于沙漠、高原等偏远地区，受限于自然条件和电力等相关基础设施，地面网络对这些地区的覆盖也很困难。

在我国领土范围外，地面网络虽然能够通过国际出口和其他国家的网络进行互联互通，但是我国对其他国家的网络并不具备控制能力。

(2) 应急响应能力不足

在出现大的自然灾害、战争时，地面网络存在着容易被摧毁，应急响应能力不足的问题。例如，2006 年 12 月 26 日晚至 27 日凌晨，中国台湾南部海域发生强烈地震，电信和网通的多条海底通信光缆均受到地震影响，发生不同程度的中断，导致国内部分用户的国际长途、国际互联网网站访问受到严重影响。经过 20 多天的紧张抢修，因地震中断的国际通信业务才全部恢复。又如，2008 年 5 月 12 日汶川地震发生后，致使中国移动 2 300 个基站受到影响，中国联通在汶川地区的 G/C 两网全部中断，中国电信在四川多地区的本地通信全面受阻。四川、甘肃、陕西三省公众通信网的累计直接经济损失达 68.8 亿元。

3.2.3 空间网络发展现状

我国航天技术飞速发展，已经形成了通信卫星和中继卫星两大系列的空间信息网络系统。

(1) 通信卫星系统

自 1984 年第一颗自主研制的通信卫星发射成功，历经近 30 年的发展，我国现拥有在轨运行的通信卫星数目已达 20 余颗，逐步形成系列。

总体来说，我国虽然已经建立起了比较完整的卫星网络体系，但是由于受技术基础限制，我国卫星无论是在质量上还是数量上与美国和俄罗斯都有很大差距，亟待追赶。通信卫星的服务范围主要是国内及周边地区，地面站也仅限于国内及少数海外站，无法满足全球覆盖的要求。同时，我国拥有的在轨运行的中星、亚太等系列卫星都属于固定通信卫星，因此，至今我国尚无自建的国内商用卫星移动通信系统，现用或准备使用的都是外商建设的全球覆盖或区域覆盖的卫星移动通信系统，如海事卫星系统、亚洲蜂窝卫星系统、铱星系统以及全球性系统，对外依存度较高。

(2) 中继卫星系统

跟踪和数据中继卫星（简称中继卫星）可为卫星、飞船等航天器提供数据中继和测控服务，极大提高各类卫星的使用效益和应急能力。世界各航天大国都在积极开展中继卫星的研发工作。目前，美国与俄罗斯两国的中继卫星系统均已进入应用阶段，并正在发展后续系统。我国迄今共发射了三颗中继卫星。三颗中继卫星形成了包括东、中、西三星组网的中继卫星系统，通过点波束实现全轨道覆盖，这一系统也是中国第一个提供全球范围实时信息传输服务的卫星系统^[16]。

数据中继卫星系统相对基于地面测控站的传统测控体系的优势表现为测控/通信覆盖率高、高度的实时性和优异的经济性。

测控/通信覆盖率高。中继卫星系统对于中低轨航天器的覆盖率很高，地基系统中一个地面站的轨道覆盖率只有 2% ~ 3%，而一颗中继卫星的覆盖率在 50% 左右。

可实时回传数据。中低轨道的航天器可以通过数据中继卫星实时回传热点地区和敏感突发事件的侦察信息，提高了反应速度，这种效果是地面测控站网络根本无法实现的。实时回传数据还加强了对航天器状态的监控能力，提高了航天器的安全性，这也有利于更好地完成实时性强的任务。

建设/维修费用低廉。中继卫星既没有陆上测控站和海上测量船等为数众多的操作人员，也没有海外陆上测控站易受政治因素影响的问题。中继卫星的使用可以取代大部分的地面测控站的任务，成为航天测控网络的主力，降低测控网的建设、运行费用。

我国的中继卫星系统利用国土在经纬度方面的跨度来东西设站，通过三颗星组网，实现了对中低轨航天器的 100% 覆盖。但因为容量有限，只能满足重要时敏数据（比如测控信息等）和部分空间信息的实时传输，大量空间信息（比如遥感图像信息等）仍然需要依靠地面站接收。

3.2.4 总结

我国地面网络得到了迅猛发展，国际互联能力不断增强，已可满足绝大部分应用需求，部分下一代互联网技术已达世界领先水平，但是地面网络在境内难以实现“全”覆盖，对境外网络无控制能力，在出现大的自然灾害、战争时，地面网络存在着容易被摧毁，应急响应能力不足的问题。

我国通信卫星的研制与应用经过 40 年的发展已取得较大的成就，但是卫星的总体性能和技术水平与欧美等强国相比还有较大的差距。在卫星总体性能和技术水平方面，国内商用通信卫星多为传统的 C、Ku 频段转发器，无法大范围地支持面向个人的移动、宽带等新业务。在通信卫星的系统设计方面，仍然采用传统的星地链路接通的基本思路，针对新业务卫星的天地系统融合式设计的理念和方法尚未完全建立、完善。同时，由于发展初期卫星资源过于分散，各自成体系、还缺少统一的中长期规划和统一管理，还缺少互联互通能力，致使我国卫星通信一致没有形成规模化的产业和完整的服务体系。

鉴于此，无论从通信卫星的研制水平，还是推动卫星应用与产业化的角度来看，建设天地一体化网络都将是未来通信卫星事业发展的主要思路。

3.3 国内外差距

通过国内外卫星系统现状分析可以看出，与国外的先进水平相比，我国的卫星系统存在以下差距：

1) 我国还没有一个能向全球提供服务的卫星通信系统，目前通信卫星的覆盖范围仅限于我国国土及周边地区；

- 2) 我国还没有一个空间组网的卫星通信系统，目前还是处于天星地站的阶段；
- 3) 我国尚无自建的卫星移动通信系统，使用的都是外商建设的移动通信系统，对外依存度较高，存在安全隐患；
- 4) 不同系统独立发展，用户主要租用卫星转发器组建专网，各专网间、卫星系统与地面网之间的互联程度低，没有达到资源和信息共享的目的，利用效率低。

另外，我国与国外也存在国情差异，不能完全照搬国外的发展经验，具体体现在：我国没有全球分布的地面站网络，建设天地一体化网络时无法采用天星地网的组网架构，因此考虑采用天网地网双骨干的天地一体网络架构。

4 体系结构

4.1 体系结构的发展现状

(1) 计算机网络体系结构的基本概念

计算机网络的体系结构是对计算机网络及其部件所完成功能的比较精确的定义，即从功能的角度描述计算机网络的结构，描述了计算机网络的各部分组成及其相互关系。

人们一般采用“层次结构”的方法来描述计算机网络，即：计算机网络提供的功能是分层次的。分层可以将庞大而复杂的问题转化为若干较小的局部问题，从而使之易于研究和处理。

在层次结构的计算机网络中，“协议”指的是同等层次中，通信双方进行信息交换时必须遵守的规则；相邻层之间都有一个“接口”（Interface），它定义了下层向上层提供的原语操作和服务。因此，在计算机网络体系结构中，对于第 N 层协议来说，它有如下特性：不知道上/下层的内部结构、独立完成某种功能、为上层提供服务、使用下层提供的服务。

(2) 互联网的 TCP/IP 体系结构

互联网的 TCP/IP 体系结构的设计与其当时的设计目标是紧密相关的。在互联网设计时，最基本、最重要的目标就是充分利用现存网络实现“网际互联”，为此设计的网络连接层（IP）负责整个互联网的全网通达，是互联网体系结构的核心^[17]。

互联网体系结构以网络层 IP 协议为核心，通过采用无连接分组交换技术提供最小的服务功能、凭借“简洁、高效、尽力而为”的特点，使得网络层具备高度适应性和可扩展性，并能够向下兼容各种通信网络技术；同时，在传输层基于 TCP 协议采用确认、重传等机制实现端到端的可靠传输，从而能够向上支持各种应用。相应地，互联网体系结构的层次模型如图 9 所示，主要包括网络接口层（数据链路层）、网络层、传输层和应用层。

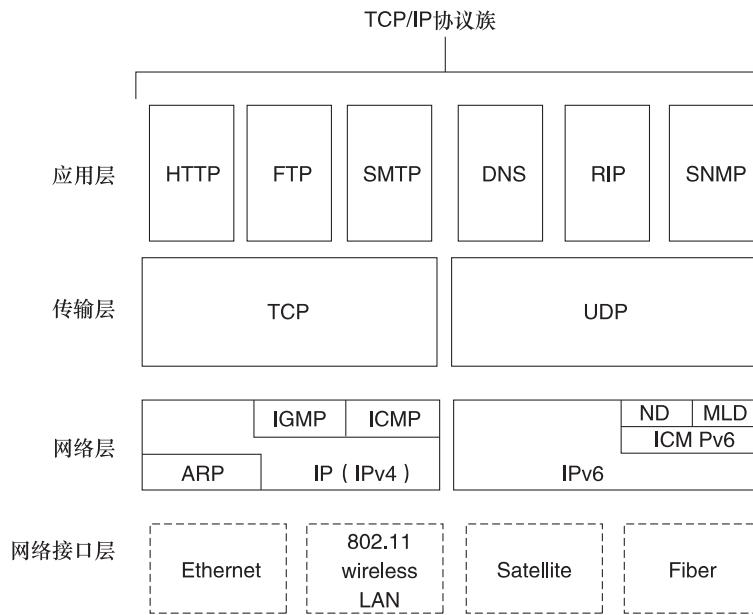


图 9 TCP/IP 体系结构层次模型

然而，随着互联网的日益普及，异构环境、普适计算、泛在联网、移动接入和海量流媒体等新应用的不断涌现，人们对互联网的规模、功能和性能等方面的需求越来越高，互联网面临着越来越严重的技术挑战，这些挑战主要包括：可扩展性问题、安全性问题、实时性问题、移动性问题、易管理性问题、可重构性问题、节能问题等。

为了解决互联网面临的技术挑战，国际互联网标准化组织 IETF 于 1998 年正式发布 IPv6 协议标准，并由此提出了下一代互联网的概念。目前，采用演进性技术路线的下一代互联网是未来的互联网发展的主流。这是因为，互联网经历了 40 多年不断演进、不断创新的发展历程，是全人类智慧的结晶，未来的互联网应该保留其具有生命力的技术优势，而且这些技术优势也一定会在支持互联网的可持续发展中发挥重要作用。与此同时，未来的互联网和非 IP 网络侧重于探索性研究，这有助于网络创新与更长期的发展。

在空间网络中采用地面互联网体系结构，其优点主要包括可以使用商用网络设备与地面互联网直接互操作实现端到端通信，技术成熟度高、能缩减航天成本、易于升级以满足未来航天任务的需要等。美国哥达德航天中心、NASA 等通过一系列的地面试验及飞行搭载试验，证明了在空间中使用地面 IP 协议的可行性，为地面互联网向空间延伸提供了技术基础^[1, 18, 19]。

(3) CCSDS 体系结构

CCSDS（国际空间数据系统咨询委员会）于 1982 年 1 月由全球主要航天组织机构联合成立，负责开发和建立适应于航天测控和空间数据传输系统的各种通信协议、数据传输规范。

该协议体系针对空间数据传输系统存在的传输距离远、节点动态性高、链路时延变化大、链路不对称、间歇性的链路连接等特点进行优化，其体系模型如图 10 所示。可以

看出，CCSDS 借鉴了开发系统互联参考模型（OSI）的分层思想，其体系结构自下而上包括物理层、数据链路层、网络层、传输层和应用层。其中，每一层又包括若干可供组合的协议，例如：SCPS-FP（文件协议）、SCPS-TP（传输协议）、SCPS-NP（网络协议）、SCPS-SP（安全协议）等^[20-28]。

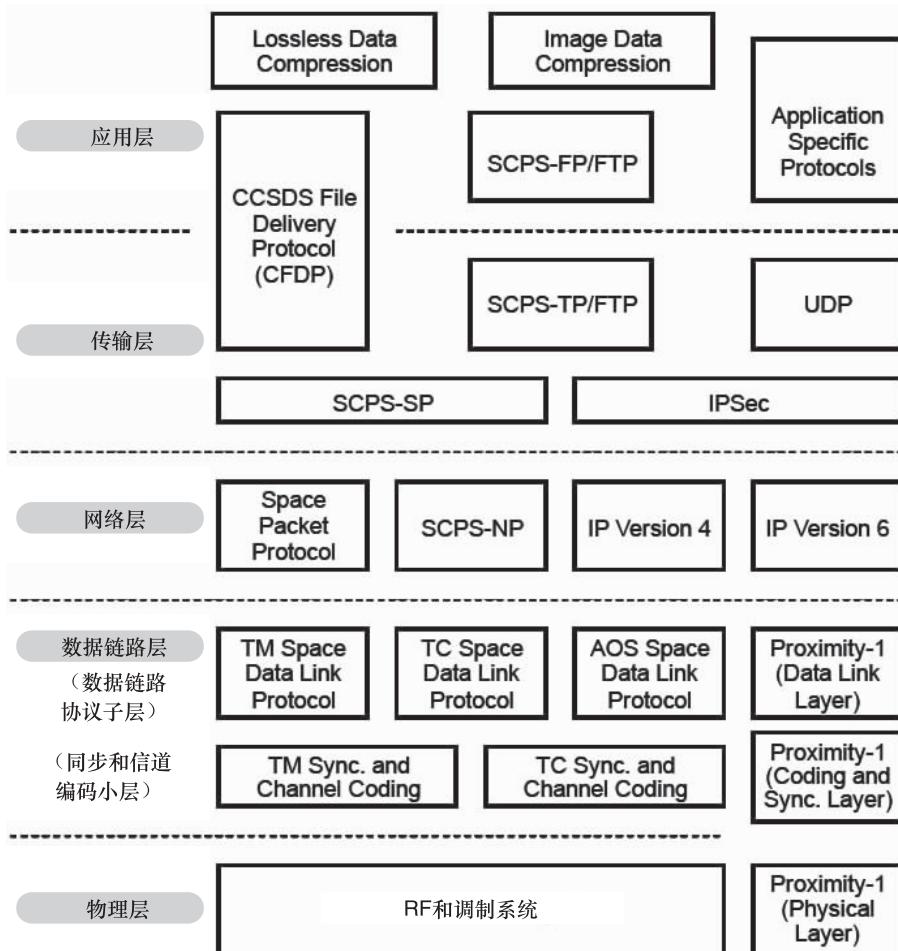


图 10 CCSDS 体系结构层次模型

CCSDS 已被较多的航天机构采纳和应用，并且已经经过了多次航天任务考验。据统计，国际上采用 CCSDS 建议的航天任务已超过 600 个。同时，为适应地面互联网的快速发展、与 TCP/IP 协议族兼容，2012 年 9 月发布了“IP over CCSDS Space Links”正式推荐标准（蓝皮书），在 CCSDS 的空间链路层协议（AOS、TC、TM、Proximity-1）上实现 IP 数据分组的传输^[29]。

CCSDS 协议体系存在的问题主要体现为：①无法与地面互联网直接互操作，需要进行协议转换；②静态路由支持能力强，但移动接入能力差，需要支持动态；③开发、测试、维护费用相对 TCP/IP 较高等。

(4) 延迟容忍网络

DTN (Delay Tolerant Networking, 延迟容忍网络) 起源于 1998 年美国国家航空航天局 (NASA) 喷气推进实验室 (Jet Propulsion Lab, JPL) 对星际互联网 (InterPlanetary Internet, IPN) 的研究。星际互联网具有不同于传统地面互联网的特点, 主要体现在: 节点之间的传播延时会非常大且变化同样很大、节点之间往往由于天体遮挡而难以保证网络的持续连通 (即间歇性连接现象) 等^[30,31]。

DTN 体系结构如图 11 所示。它在应用层之下传输层之上引入了一个 Bundle 层 (覆盖层), 并通过使用磁盘等永久存储方式来克服网络的间歇性连接问题。Bundle 层提供了类似于网关的功能, 可以在各个底层协议之间 (如 TCP/IP、CCSDS 等) 提供互操作。

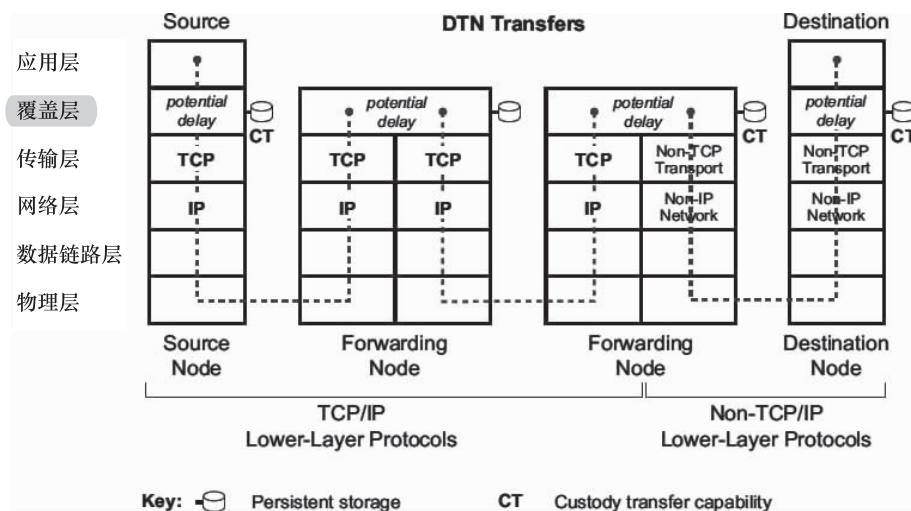


图 11 DTN 体系结构的层次模型

NASA 通过太空 DTN 发展计划, 进行了一系列的 DTN 试验, 应用领域涵盖了遥感、深空探测、空间站、中继通信等领域。实验探索了 DTN 在飞行实验中可能遇到的问题, 对 DTN 技术的自动存储转发机制、网络协议和软件实现等关键技术要素进行了探索和验证。实验表明, DTN 能够正常运行在长时延、有中断情况, 以及非对称链路和单向链路上^[32-36]。

(5) 比较和总结

TCP/IP、CCSDS 和 DTN 体系结构的简要对比如表 4 所示。

表 4 现有主要网络体系结构对比

协议体系	主要特点	优势	问题
TCP/IP	核心技术 IP: 无连接分组交换; “窄腰”结构	技术成熟度高、研发成本低; 天基网和地面网直接互操作, 实现“一体化”	不适应空间链路特点 (大延时、断续连接、高误码率等); 自身存在安全性等挑战
CCSDS	专为空间链路设计	协议体系完善; 经过多次航天任务考验, 较多航天机构已采纳	无法与地面网直接互操作, 存在协议转换; 开发维护费用较高

(续)

协议体系	主要特点	优势	问题
DTN	“覆盖层”方式，灵活性高；存储转发机制	面向星际互联网设计；兼容性较好，可基于已有的成熟网络技术	还处于研究阶段，标准的制定刚刚起步

通过调研分析和比较，我们在网络体系结构设计方面得到了如下结果：

1) 主流网络体系结构 (TCP/IP、CCSDS、DTN) 均采用分层思想，它们的分层方式基本相同，主要都包括物理层、数据链路层、网络层、传输层、(DTN 覆盖层)、应用层等协议层次。

2) IP 技术已获普遍认可，基于 IP 技术可实现不同网络的互联。

面对空间网络的互联以及空间网络和互联网的互联，从技术发展趋势来看，IP 协议的无连接存储转发方式、IP 分组格式等已经获得普遍认可。

3) 通过网关转换等方式基本可实现不同网络的互通。

虽然在 IP 层之上，不同的体系结构采用了不同的传输层协议 (如 TCP、UDP、SCSP-TP 等)，但是通过网关转化，可以实现端到端的数据传输，以及可选择的端到端可靠数据传输。

4) 现有的互联互通方式的效率低、效果差，尚未形成有效的天地“一体化”网络体系。

虽然通过 IP 技术能实现异构网络的互联，通过网关转化方式能实现异构网络间数据的互通 (端到端传输)，但是目前的互联互通方式效率比较低下，尚未形成有效的天地“一体化”网络体系，例如尚未形成有效的一体化网络路由等。因此，面对天地“一体化”网络体系架构，还需要重点突破“编址和路由”、“端到端传输”等关键技术，形成有效的一体化路由体系和端到端传输体系。

4.2 体系结构的研究难点

天地一体化网络规模庞大、结构复杂、支持的业务种类繁多、网络伸缩性强、时空跨度大、网络拓扑结构不断变化，并且涉及互联网、空间网络等多种异质异构网络。用传统的网络理论、技术方案难以建立反映空间网络动态时空关系的网络模型和解决其中的关键技术问题。目前，这些网络具有不同的体系结构和网络结构，如何对异构网络进行有效融合也成为研究的难题。而且，空间网络本身还呈现“烟囱式”的发展模式，不同卫星系统相对独立、专用、缺乏统一的网络协议规范。因此，要想将这些网络进行有效整合、实现深入融合，还面临着众多的难题。

如何把复杂多样、规模巨大的一体化网络抽象为结构清晰、功能简捷、易于实现的网络体系结构，在体系结构的框架指导下构造网络的各组成部分并建立相互关系，使网络既能适应通信技术的快速发展与变化，又能支持层出不穷的新型应用，这对一体化网络体系结构研究提出了挑战。

4.3 体系结构的发展趋势

(1) 坚持分层的原则

分层对网络体系结构至关重要，基于此，可以将庞大而复杂的问题简单化。同时，当前适应天地一体化信息不同场景的3种主流网络体系结构均采用基本相同的分层方式。一体化网络规模庞大、异质异构性强，并包含3种主流协议的适应场景，使得分层成为体系结构必不可少的设计原则。

(2) 坚持IP技术精髓，有效吸纳其他研究成果

互联网体系结构以网络层IP协议为核心，通过采用无连接分组交换技术，凭借“简洁、高效、尽力而为”的特点，使得网络层具备高度适应性和可扩展性，能够向下兼容各种通信网络技术，从而促进了互联网在全球范围的极大发展。

目前，尽管IP协议主宰着连接到卫星用户终端的地面网络系统，但大部分的卫星系统仍采用ATM作为链路层技术来连接各卫星用户终端。这主要是因为这些系统技术本身与ATM相契合，而且在系统设计之初，ATM被视为是未来主流的网络技术。纵观国外的典型系统，它们也都在研究卫星网络中的IP路由问题。例如，虽然Teledesic系统在卫星间的链路、空地间的链路都采用自己设计的协议，Spaceway和Astrolink采用基于ATM的星间链路和空地链路以及自定义的MAC/LLC协议，Skybridge采用ATM作为地面部分协议，但这些系统都通过开辟通道来支持IP协议。

因此，在天地一体化网络中应继续坚持IP协议，并以此为基础针对一体化网络的特点开展体系结构的研究。同时，还应以互联网体系结构为基础，有效吸纳CCSDS体系结构和DTN协议体系面向空间网络的研究成果，取长补短。

5 组网架构

5.1 网络结构的发展现状

信息网络的结构反映了构成网络的“节点”与“链路”之间的拓扑关系，不同的网络结构将使得信息网络具有不同的基本特性。

(1) 互联网的网络结构

互联网是一批独立自治的计算机系统的互连集合体，包括资源子网（或用户子网）和通信子网两部分。资源子网由服务器、客户计算机等终端设备组成；通信子网则由各种网络互连设备（路由器、交换机、HUB等）及其通信线路组成。互联网通常分为主干网和接入网（如企业网、园区网等）两级结构。

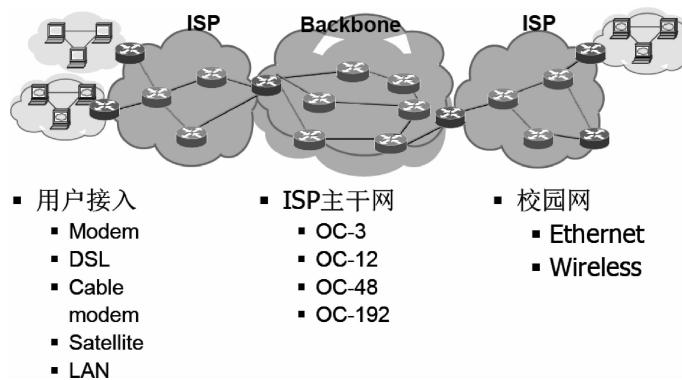


图 12 互联网的网络结构示意图

(2) 天基网络的网络结构

目前天基网络系统主要有两大类：地球静止轨道（GEO）卫星系统、中低轨（LEO）卫星系统，这两种卫星系统的比较如表 5 所示。可以看到，虽然从单个制造和发射成本考虑，GEO 卫星比 LEO 要高，但是 LEO 系统需要的卫星多于 GEO 系统，因此总成本可能是 GEO 系统低。从通信质量考虑，LEO 卫星系统最有优势，其信号传输衰耗和延迟都是最低的，用现有工艺可以小型化。另外，由于 LEO 卫星系统需要的卫星数量较多，整个网络的技术复杂度较高，系统建设和维护的难度较大。经过综合考虑，建议我国采用基于 GEO 卫星的天基网络系统。

表 5 不同轨道通信卫星系统的比较

卫星系统名称	GEO	LEO
需要卫星数量	少	多
单个卫星成本	最高	最低
卫星寿命（年）	15 ~ 20	5 ~ 8
网络复杂性	低	高
手持移动终端	很困难	可能
传播延时	大	小

(3) 天地一体化网络的网络结构

经过调研分析，我们可以将天地一体化网络的结构归为 3 大类：天星地网、天基网络、天网地网，不同网络结构的比较如表 6 所示。

表 6 不同天地一体化网络结构的比较

网络结构	天星地网	天基网络	天网地网
地面网络	全球分布地面站网络	系统可不依赖地面网络而独立运行	天地配合，地面网络不需要全球布站
星间组网	否	是	是
星上设备	简单	复杂	中等
系统可维护性	好	差	中
技术复杂度	低	高	中
建设成本	低	高	中

1) 天星地网。天星地网是目前普遍采用的一种网络结构，包括 Inmarsat、Intelsat、WGS 等系统，其特点是天上的卫星之间不组网，而是通过全球分布的地面站实现整个系统的全球服务能力。在这种网络结构中，卫星只是透明的转发通道，大部分的处理还是在地面完成，因此星上设备比较简单，系统建设的技术复杂度低，升级维护也比较方便。

2) 天基网络。天基网络是另一种网络结构，典型的系统有 Iridium、AEHF 等，其特点是采用星间组网的方式构成独立的天基网络，整个系统可以不依赖地面网络而独立运行。这种网络结构弱化了对地面网络的要求，把处理、交换、网络控制等功能都放在星上完成，提高了系统的抗毁能力，但由此也会使星上设备复杂化，导致整个系统的建设和维护成本较高。这种单纯的天基网络结构从商业上来说并不算成功，它主要是基于军事上对网络极端抗毁性的需求而设计的。

3) 天网地网。天网地网介于上述两种网络结构之间，以 TSAT 计划为典型，其特点是天基和地面两张网络相互配合共同构成天地一体化网络。在这种网络结构下，天基网络利用其高、远、广的优势实现全球覆盖，地面网络可以不用全球布站，但能完成大部分的网络管理和控制任务，简化整个系统的技术复杂度。

综合考虑之后，我们认为天网地网是适合我国国情的天地一体化网络的结构。

5.2 组网架构的研究难点

(1) 多层次协作的网络组网架构总体设计

未来一体化网络的成员节点种类繁多，不同网络链路的传输特性各异，各节点的功能及组网特性不同，如何进行协同设计、发挥综合效能，是一个很大的难点，当前还没有现成的网络拓扑结构可供借鉴。如今，各独立网络已经有比较成熟的组网架构，但涵盖高低轨卫星、临近空间平台、具有地面移动通信网络节点以及固定互联网节点等多层次、综合应用的一体化网络的架构还停留在概念阶段。

(2) 多层次异构网络的有效互联和融合

一体化网络的组成复杂，不同组成部分在网络拓扑稳定性、网络规模等方面差异巨大，其相互之间的连接方式以及由此形成的组网结构，将对一体化网络整体的互联和融合效果带来较大的影响。如果空间网络的任意设备（如卫星或飞行器）都可以通过地面关口站直接与地面互联网主干网相连并进行路由交互，那么空间网络的高度动态性将会引发地面互联网主干网的路由振荡。因此，一体化网络组网架构的设计和规划与编址、路由、端到端传输等网络关键技术是密切相关的，如何对这些研究内容进行统筹考虑和联合设计，是多层次异构网络实现一体化有效互联和融合的难点。

(3) 空间网络接入资源的统一规划

空间网络接入资源具有多层次、多维度特点，在空间上包括 GEO 卫星资源、LEO 卫星资源、临近空间资源；在频率上可能涵盖 S、Ku、Ka 频段以及光等；其他还有时隙、功率等资源。空间网络的通信拓扑关系和时空关系复杂，空间接入资源的高效利用和各类具备不同能力的用户的需求满足是空间网络接入资源统一规划的难点。目前各类独立

的空间网络已经能够对其系统内的接入资源进行优化分配，并具备资源管理控制的能力，但要从空间一体化的角度对整体接入资源进行统一规划还会面临较大的挑战。

5.3 组网架构的发展趋势

由于空间应用成本等原因造成星载、机载、舰载等网络节点功能有限，此外的节点同普通有线网中的节点相比有着很大的局限性。同时，应用子网由于用途不同，数据传输需求量差距很大，所以难以用统一网络覆盖。因此在未来根据空间资源以及应用需求，采用具备星-地、星间通信能力的空间移动通信系统作为高数据率传输的主干网，且各类卫星（星座）、地面固定、车载、机载、舰载等应用子网接入的方式来构成应用信息系统。

目前，国外（尤其是美国）主要采用“天星地网”的方式来构建空天宽带网络，各种子系统之间的连接主要在地面完成。但是由于国情不同，我国地面站的全球布站在可行性上面临着极大的挑战，所以完全依赖地面站的方式并不适合我国。根据我国国情和发展需求，需要采用“天网地网”的方式建设未来一体化网络。

6 关键技术

天地一体化网络涵盖了空间网络、地面移动网络、地面互联网等多种网络环境，其所具有的大时空跨度、拓扑高动态、网络立体化等特点，对现有网络的路由、传输以及星上信息处理等关键技术提出了巨大的挑战。

6.1 路由技术

6.1.1 空间网络路由和编址技术

空间网络独有的技术特点使得传统互联网的路由技术无法适用于空间网络环境，因此必须要进行新型网络路由协议的系统设计和实现。和传统的地面互联网相比，空间网络的主要特点包括：①空间网络组成复杂，结构立体化；②空间网络的节点高速运动，网络拓扑结构时变性强；③空间网络的通信环境恶劣，链路质量差，误码率高；④空间网络的通信距离远，通信时延及时延抖动高；⑤空间网络载荷有限，能源有限，节点计算、存储和带宽资源受限；⑥基于空间网络任务的多样性，对空间网络的传输需求也具有多样性；⑦空间网络节点之间的通信方式开放，安全性能差。

针对空间网络节点的地址编制方案，可以将相关地理位置信息和节点物理标识综合考虑，空间节点在通过不同的地面网络节点（关口站）和互联网通信时，使用不同的网络地址，这样可以避免由于节点移动性带来的路由稳定性问题。当然，如此，就必须要有

考虑用相关的机制来维护空间节点的物理标识与地址之间的映射关系。

同时，空间节点的运动速度虽然较快，但是其轨道相对固定，运动位置具有可预测性和周期性，因此可以通过节点轨道计算进行预测，并和一小部分的实际测量信息相结合，构造空间网络节点拓扑结构数据库，并集中进行路由信息计算，形成路由表。根据空间网络的运行特点，给路由表增加时间维度，例如路由有效期，以及失效备选路由等，形成时变路由策略，同时可以结合空间节点的应用特点，由地面中心控制节点统一计算路由信息，再上传至空间节点。

6.1.2 空间网络与地面网络的路由融合技术

空间网络的覆盖范围广（海洋、天空等），同时网络较为复杂，卫星计算及存储资源极为匮乏；地面网络覆盖范围有限，同时较为稳定，计算级存储能力充足。如何融合地面网络和空间网络的优势，实现全面可靠的“空-空”、“空-地”、“地-地”通信是一体化网络路由融合需要解决的关键问题。

空间网络和地面网络的路由深度融合主要涉及 5 个层次：地面互联网、地面站系统、非静止轨道卫星网络、静止轨道卫星网络和临近空间网络。可采用软件定义网络（Software Defined Networking，SDN）模型实现一体化网络中的多网融合。

SDN 的核心思想是集中式控制，即多个交换设备连接同一个控制器，控制器通过中心化的方式做出决策，并将路由策略下载到交换机，从而实现高效的路由转发。在一体化网络中，空间网络设备由于成本等原因，在计算能力、存储能力等方面极为匮乏，无法支撑当前互联网中的动态路由协议。而 SDN 弱化了一般网络设备功能，只对集中式控制器有更强的需求，这在一定程度上契合空间网络设备的特点。同时，空间网络需要必要的分布式策略，保证在星地链路失效时，空间网络有自愈机制。

6.1.3 空间路由器的关键技术

空间网络相对于地面网络，面临资源受限、链路状态多样性等问题，使得传统地面的路由器技术不能直接应用在星上。

此外，受限于国外器件禁运及国内星载设备的处理能力，星载 IP 路由器的软硬件平台只能在开放操作系统软件、国产 CPU 处理器及中小规模的 FPGA 等范围内选择，面临着软硬件兼容性、功能模块高效设计等众多挑战，因此星载 IP 路由器的软硬件平台设计实现是主要的技术难点。

6.2 传输技术

6.2.1 端到端传输技术

DTN 和 CCSDS 协议体系是面向空间网络的两大协议体系。正如 4.1 节中所述，DTN 主要用于时延长、容易出现链路中断的网络环境中。DTN 在应用层之下传输层之上引入

了一个 Bundle 层（覆盖层），并通过使用持久存储方式来克服网络的间歇性连接问题。DTN 的持久存储将不同时间维度的端到端传输联系到了一起。

面对空间通信中的高误码、断续、大时延、不对称等链路特点，CCSDS 开发了 SCPS-TP 空间可靠通信协议。SCPS-TP 支持可选的默认丢包的原因假设（误码或网络拥塞）、可选的拥塞控制机制，以及对误码、网络拥塞、链路中断的显式判断机制来优化空间网络端到端传输的拥塞控制机制；通过窗口缩放和修改定时器来减少长往返时延对数据传输的影响；针对空间链路带宽有限的限制，SCPS-TP 通过头部压缩和 SNACK 技术来提高数据传输吞吐量；采用降低确认应答频率以及进行数据传输速率的控制来适应空间链路不对称的特点。

随着端到端传输技术的不断发展，各种针对端到端传输协议的优化方案被提出。针对空间通信中的高误码、断续、大时延、不对称等链路特点的相关改进方案具体如下。

针对大时延特性的相关研究而言，仅修改从终端角度出发的方案的主要思路是提高窗口增长的速度，包括 TCP hybla^[39] 和在 Linux 内核中广泛使用的 TCP cubic^[40]。需要修改路由器的方案主要包括：TCP peach^[41] 及 TCP peach +^[42]，提出在慢启动、快速恢复阶段发送低优先级数据包，网络发生拥塞时优先丢弃低优先级数据包。在网络不发生拥塞时，低优先级数据包到达接收方后加快发送方窗口的增长；P-XCP^[43] 中的路由器根据其队列中竞争的数据流计算出各个流的合理带宽，通过和发送方的协同决定发送速率；ATSP^[44] 在路由器中采用主动队列管理 AQM，按照指定的队列长度提前进行数据包的标记和丢弃。

现有的针对误码的 TCP 优化机制主要包括：第一类是采用分裂连接的方法，将端到端的连接分成多段，使得空间网络的丢包行为不影响地面网络的发送，如 TCP PETRA^[45]。第二类采用的是区分误码造成的丢包和拥塞造成的丢包的方式，如 TCP westwood^[46]/westwood + 采用被动测量方式获得链路的带宽，在发生丢包时以获得的链路带宽进行窗口的变化，而不是直接减半。第三类研究采用的是冗余编码的方式。CTCP (Network Coded TCP)^[47] 提出在卫星网络中将网络编码与端到端传输相结合，并采用线性编码方式对发送的数据包进行处理，在网络发生少量丢包时通过编码进行数据包的恢复，提高端到端传输的吞吐率。

针对非对称特性带来的 TCP 性能下降问题，现有方案主要对采用选择确认和降低确认频率的方式进行优化。STP 和 SCPS-TP 通过采用选择确认和降低确认频率，使得多个数据包对应一个确认包，从而降低所需的上下行带宽比。AF (ACK Filtering) 在路由器处对 ACK 进行合并和过滤，由于 ACK 是累计确认的，所以合并后反向路径需要发送的包随之减少，同时发送窗口仍然能够继续增长。

低轨卫星间的周期运动使得端到端的传输参数随着时间周期动态地发生变化。TCPW-BLC^[48] 针对 Iridium 网络端到端传输时延的动态周期变化，提出在 TCP westwood + 拥塞控制算法基础上，利用可以预测的端到端传输时延重新计算 TCP 的超时时间并更新 westwood + 的估计带宽。

从对现有的研究总结可以看出，目前相关研究主要侧重于对端到端传输空间链路的

某方面特性进行优化，尚未提出整体的传送层解决方案。

与此同时，天地一体化网络涉及地面网络、空间网络等多种异构网络，自然地为端到端传输提供了多种可选途径，因此我们需要研究多种异构网络的高效协同传输机制，从而提升一体化网络端到端传输的整体容量和适变能力。但是，由于不同网络在覆盖范围、带宽、时延、动态性等传输特性和性能方面有着巨大的差异，使得天地协同传输面临着新的难题、挑战。

因此，针对互联网、空间网络、无线移动网络的传输资源、能力和差异性，需要进行统筹考虑，并通过端到端多网络一体化传输架构及关键技术实现异构网络传输资源的联合管理和协同优化。

同时，SDN 的集中控制机制为网络状态信息的获取提供了方便，也为异质异构网络间的协同提供了途径，并且 SDN 的控制与转发分离机制使得数据分组的网络传输路径能够以近乎实时的方式进行修改，为一体化网络中的动态路径选择、调度和分配提供了基础支持。

6.2.2 链路传输技术

目前，国际上已完成了空间激光通信链路的概念研究，其中的关键技术和核心部件已解决，并实现了低轨卫星对同步卫星的低、中码速率激光通信实验，进行了低轨卫星对地面站的激光通信实验。这些通信实验中的系统达到了高捕获概率，短捕获时间，抗多种干扰的高灵敏度动态跟踪和较高的传输数据率，同时研制了激光链路系统评估测试平台及分析、仿真软件。

国内高速率空间信息传输技术已实现 1.5Gbps 微波传输和 2.5Gbps 激光通信，正在向微波 3Gbps、激光 5~10Gbps 方向发展。目前星间仍以微波传输为主，对于激光通信，我国已有 20 多个单位开展了相关研究，并取得了重要成果。哈尔滨工业大学、长春理工大学等已经完成了星地间激光通信和用飞机模拟的星间通信的试验验证，并取得不错的成果。

但是，总体而言，我国对空间网络的高动态、高速率信息传输还没有突破瓶颈，微波通信也未能实现空间信息的宽带传输，激光通信尚未实用化，量子通信、太赫兹通信仍处于实验研究阶段。

6.3 星上信息处理技术

空间信息融合处理技术主要有加权平均法、卡尔曼滤波、贝叶斯估计、统计决策理论、D-S 证据理论、模糊逻辑、神经网络、粗糙集理论等。

在信息快速提取和处理方面，美国已实现天基平台从获取信息到快速反应仅用 1~3 分钟。

在信息尺度/粒度、信息相容/相悖、完整/不完整测量、时序/非时序测量、可观测性强/弱等特征上具有较大差异的信息融合问题在实际融合过程中经常出现，这一问题也是当前信息融合领域的前沿问题，至今还没有成熟的基础理论和方法与之应对。另外，

多平台图像配准、多介质图像特征关联、多粒度（多分辨率）图像稀疏变换、多源图像目标融合检测与识别等技术研究至今仍处于起始阶段，目前还存在诸多难点尚未解决。

从天基信息获取到反应时间境内要 10 小时、境外要 30 小时，对空间信息稀疏表征、大数据处理和信息融合尚未形成完整的理论体系，在异步、异类多传感器信息的融合、海量信息的处理等方面还有待突破，这就是目前我国的一体化网络现状。

当前一体化网络面临 3 个多元化：①网络类型多元化，包括互联网、无线移动网和空间网络在内。②通信场景多元化，涵盖深空通信、低空通信、地面通信、水域通信等不同通信场景。③网络需求多元化，需要满足商用、政府、军用和个人通信等不同场景下的网络需求。

此外，由于星上载荷受限等造成的通信系统内存小、CPU 处理能力低、所带电源有限……，使得天地一体化网络中的节点与普通有线网中的节点相比有着很大的局限性，增加了星上处理和系统实现的难度。

7 结束语

以美国为代表的发达国家虽然近些年来系统地发展了窄带、宽带、受保护通信卫星及数据中继卫星系统，并开始快速推进临近空间网络，但是这些空间信息系统的互联互通互操作能力仍然较差，因此按全球信息栅格的理论来建立统一空间信息系统是美国未来天地一体化网络的长远发展趋势。

我国的卫星网络体系建设起步较晚，和国际高水平国家相比还有较大差距。目前我国是按照任务规划卫星系统，空间信息系统呈现“烟囱式”的发展模式，各系统相对独立、专用。这种自成体系、条块分割的现象，导致信息共享能力差、多源信息难以融合和综合利用、重复建设严重等问题，而且星上、星地和地面三部分网络协议体系尚未进行一体化设计，缺乏统一的网络协议规范，从而使得各系统之间难以有效地互联互通，也难以实现资源和信息共享，利用效率低。根据我们对空间协议的调研研究，我国天地一体化网络的协议体系以 IP 技术为基础，实现网络间互联互通，并融合 DTN、CCSDS 等协议机制。结合我国没有全球分布的地面站网络的国情，建设我国的天地一体化网络不能完全照搬国外的发展经验，也不能采用国外的天星地网组网架构，因此我国的天地一体网络需要采用天网地网双主干的网络架构。

天地一体化网络的大时空跨度、拓扑高动态、网络立体化等特点，为一体化网络路由、传输、星上信息处理等关键技术带来巨大挑战。本章围绕天地一体化网络，针对国内外相关卫星系统展开调研，总结国内外卫星系统的发展现状，并通过对国内外一体化组网、传输、星上信息处理等关键技术的调研分析，给出我国天地一体化网络的发展趋势，为我国天地一体化网络的相关研究提供参考。

注：本报告的内容主要来自工业和信息化部电子科学技术委员会 2014 年重点课题“天空地一体化信息网络的体系架构研究”的项目研究报告。

参考文献

- [1] 沈荣骏. 我国天地一体化航天互联网构想 [J]. 中国工程科学, 2006, 8(10) : 19-30.
- [2] 李德仁. 论新一代空间信息网络—基础理论与关键技术 [R]. “天地一体化网络”高峰论坛, 2013.
- [3] 张军. 面向未来的空天地一体化网络技术 [N]. 中国航空报, 2009-04-07(3).
- [4] 张军. 天基移动通信网络 [M]. 北京: 国防工业出版社, 2011.
- [5] 吴曼青. 构建天地一体化网络建设强大的信息系统国度 [R]. “天地一体化网络”高峰论坛, 2013.
- [6] 吴建平. 天地一体化网络现状与发展思考 [R]. “天地一体化网络”高峰论坛, 2013.
- [7] 陆建华. 空间信息网络: 机遇与挑战 [R]. “天地一体化网络”高峰论坛, 2013.
- [8] <http://www.lwgw.com/NewsShow.aspx?newsId=32873>.
- [9] 史西斌, 等. 国外空天宽带网络发展现状及体系结构分析 [R]. “天地一体化网络”高峰论坛, 2013.
- [10] 潘成胜, 等. 空间信息网络的内涵 [R]. “天地一体化网络”高峰论坛, 2013.
- [11] 胡源, 等. 天地一体化网络国外发展现状与趋势 [R]. “天地一体化网络”高峰论坛, 2013.
- [12] 张忠强. Inmarsat 海事卫星通信的发展 [J]. 中国新通信, 2013, 15(15) : 27-29.
- [13] 林飞, 祝彬, 陈萱. 美国“宽带全球卫星通信系统” [J]. 中国航天, 2013, 12: 10(10).
- [14] 张紫. 第 33 次中国互联网络发展状况统计报告 [J]. 计算机与网络, 2014, 40(2) : 5-5.
- [15] <http://www.cngi.cn/>.
- [16] 陈锡明, 等. 天地一体化网络建设设想 [R]. “天地一体化网络”高峰论坛, 2013.
- [17] CLARKE, D. The design principles of the DARPA Internet protocols [J]. ACM SIGCOMM Computer Communications Review (CCR) 25, 1995, 1(1).
- [18] Scott K L, Burleigh S. Bundle protocol specification [S].
- [19] Burleigh S, Ramadas M, Farrell S. Licklider Transmission Protocol- Motivation [J]. IETF Request for Comments RFC, 2008, 5325.
- [20] 叶晓国, 肖甫, 孙力娟, 等. SCPS/CCSDS 协议研究与性能分析 [J]. 计算机工程与应用, 2009, 45(4) : 34-37.
- [21] 蒋迎春. CCSDS 空间数据通信协议研究及其 OPNET 仿真 [D]. 国防科学技术大学, 2005.
- [22] Overview of space communications proto cols, report concerning space data system standards [R]. CCSDS130.0. G.2 . 2007.
- [23] Proximity-1 Space Link Protocol—Data Link Layer [S]. Recommendation for Space Data System Standards, CCSDS 211.0-B-4. Blue Book. Issue 4. Washington, D. C. : CCSDS. July, 2006.
- [24] TM Space data Link Protocol [S]. Recommendation for Space Data System Standards. CCSDS132.0. B. 1. 2003.
- [25] TC Space Data Link Protocol [S]. Recommendation for Space Data System Standards. CCSDS232.0. B. 1. 2003.
- [26] AOS Space Data Link Protocol [S]. Recommendation for Space Data System Standards. CCSDS732.0. B. 2. 2006.
- [27] Space Communication Space Packet Protocol [S]. Recommendation for Space Data System Standards, CCSDS 133.0-B-1. Blue Book. Issue 1. Washington, D. C. : CCSDS, September ,2003.

- [28] Space Protocol Specification (SCPS) — Network Protocol (SCPS-NP) [S]. Recommendation for Space Data System Standards. CCSDS713.0.B.1.1999.
- [29] IP Over CCSDS Space Links [S]. Draft Recommendation for Space Data System Practices. CCSDS 702.1-R-2. Red Book. Issue 2. Washington, DC.: CCSDS, January, 2007.
- [30] Forrest Warthman. Delay-Tolerant Networks (DTNS): A Primer, v1.0 [R]. February, 2002.
- [31] Forrest Warthman. Delay-Tolerant Networks (DTNS): A Tutorial, v1.0 [R]. February, 2002.
- [32] Ivancic W, Eddy W M, Stewart D, et al. Experience with Delay-Tolerant Networking from Orbit [J]. International Journal Of Satellite Communications And Networking. 2010(28): 335-351.
- [33] Wyatt J, Burleigh S, Jones R. Disruption Tolerant Networking Flight Validation Experiment on NASA's EPOXI Mission [C]. The First International Conference on Advances in Satellite and Space Communications (SPACOMM 2009). Colmar, France: 2009.
- [34] Jones R M. Disruption Tolerant Network Technology Flight Validation Report DINET [R]. NASA JPL, 2009.
- [35] Davis F A, Marquart J K, Menke G. Benefits of Delay Tolerant Networking for Earth Science [C]. Proceeding of 2012 IEEE Aerospace Conference. Big Sky, MT: 2012: 1-11.
- [36] Kraft R. NASA. ESA use experimental interplanetary internet to test robot form International Space Station [EB/OL]. [2012]. http://www.nasa.gov/home/hqnews/2012/now/HQ_12-391_DTN.html.
- [37] 清华大学, 等. 一体化网络发展思路研究 [R]. 电子科技委课题, 2013-2014.
- [38] 清华大学, 等. 地一体化网络试验验证系统的现状、趋势和方案研究 [R]. 电子科技委课题, 2014-2015.
- [39] Carlo Caini, Rosario Firrincieli. TCP Hybla: a TCP enhancement for heterogeneous networks [J]. Int. J. Satellite Communications Networking 22(5): 547-566 (2004).
- [40] Ha Sangtae, Injong Rhee, Lisong Xu. CUBIC: a new TCP-friendly high-speed TCP variant [C]. ACM SIGOPS Operating Systems Review 42.5 (2008): 64-74.
- [41] Akyildiz, Ian F, Giacomo Morabito, Sergio Palazzo. TCP-Peach: a new congestion control scheme for satellite IP networks [J]. IEEE/ACM Transactions on Networking (ToN) 9.3 (2001): 307-321.
- [42] Akyildiz, Ian F, Xin Zhang, Jian Fang. TCP-Peach+: Enhancement of TCP-Peach for satellite IP networks [J]. Communications Letters, IEEE 6.7 (2002): 303-305.
- [43] Zhou Kaiyu, Kwan L Yeung, Victor OK Li. P-XCP: a transport layer protocol for satellite IP networks [C]. Global Telecommunications Conference, 2004. GLOBECOM04. IEEE. Vol. 5. IEEE, 2004.
- [44] Muhammad, Firat Kasmis, Tomaso De Cola. Advanced transport satellite protocol [C]. Global Communications Conference (GLOBECOM), 2012. IEEE, 20.
- [45] Marchese, Mario, Michele Rossi, Giacomo Morabito. PETRA: Performance enhancing transport architecture for satellite communications [J]. Selected Areas in Communications, IEEE Journal on 22.2 (2004): 320-332.
- [46] Casetti, Claudio, et al. TCP Westwood: end-to-end congestion control for wired/wireless networks [J]. Wireless Networks 8.5 (2002): 467-479.
- [47] Speidel U, Cocker E, Vingelmann P, Heide J, Médard M (2015, June). Can network coding bridge the digital divide in the Pacific? [C]. In Network Coding (NetCod), 2015 International Symposium on (pp. 86-90). IEEE.
- [48] 杨力, 李静森, 魏德斌, 等. 一种高动态卫星网络的拥塞控制算法 [J]. 宇航学报, 35.8 (2014).

作者简介

清华大学

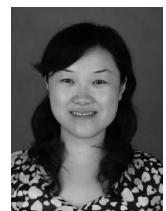
李贺武 (1974—), 男, 博士, 副研究员, 主要研究领域为无线移动网络、天地一体化网络。

E-mail: lihewu@ cernet. edu. cn。



吴 茜 (1978—), 女, 博士, 副研究员, 主要研究领域为计算机网络体系结构、无线移动网络、天地一体化网络。

E-mail: wuqian@ cernet. edu. cn。



吴建平 (1953—), 男, 博士, 教授, 博士生导师, CCF 会士, 主要研究领域为计算机网络体系结构、网络协议测试。

E-mail: jianping@ cernet. edu. cn。



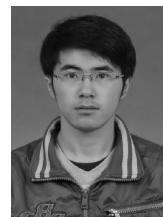
徐明伟 (1971—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为计算机网络体系结构、大规模路由。

E-mail: xumw@ tsinghua. edu. cn。



杨增印 (1988—), 男, 博士研究生, 主要研究领域为天地一体化网络、一体化路由。

E-mail: yang-zy14@mails.tsinghua.edu.cn。



江 卓 (1991—), 男, 博士研究生, 主要研究领域为天地一体化网络、端到端多路径传输。

E-mail: jiangzhuo_cs@126.com。



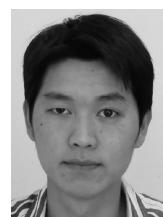
中国电子科学研究院

陆 洲 (1970—), 男, 硕士, 研究员级高工, 主要研究领域为卫星通信与网络。

E-mail: cetc_luzhou@163.com。



张 平 (1981—), 博士, 主要研究方向是空间网络体系架构和网络模型。



秦智超 (1981—), 男, 博士, 主要研究领域为天地一体化信息网络。

E-mail: qzc0308@163.com。



长春理工大学

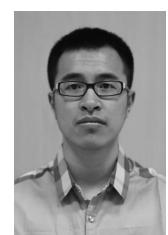
姜会林 (1945—), 男, 博士, 教授, 博士生导师, 长春理工大学学术委员会主任, 主要研究领域为光学系统设计、光电动态检测、空间激光通信。

E-mail: hljiang@ cust. edu. cn。



刘显著 (1989—), 男, 博士研究生, 主要研究方向为激光通信、相控阵通信、天地一体化网络等。

E-mail: liuxianzhu@ cust. edu. cn。



底晓强 (1978—), 男, 博士, 副教授, 博士生导师, CCF 会员, 主要研究领域为天地一体化网络安全、服务质量。

E-mail: dixiaoqiang@ cust. edu. cn。



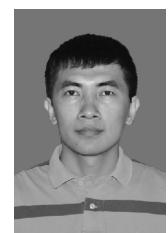
祁晖 (1983—), 男, 博士, 讲师, 主要研究领域为访问控制技术、移动通信与车联网。

E-mail: qh_odd@ 126. com。



从立钢 (1983—), 男, 硕士, 讲师, CCF 会员, 主要研究领域为计算机网络、信息安全。

E-mail: congligang163@ 163. com。



王天枢 (1975—), 男, 博士, 教授, 博士生导师, 主要研究领域为激光通信、全光网络、光纤激光器等。

E-mail: wangts@ cust. edu. cn。



“互联网 +” 战略的研究进展与趋势

CCF Petri 网专委会

摘要

随着互联网深入应用，特别是以移动技术为代表的普适计算、泛在网络的发展，互联网像水电一样逐渐成为人们生产生活、经济社会发展各行各业所必需的生产要素。“互联网 +”的思维模式已经逐步成为一种重新审视整个商业生态的思考方式。当前“互联网 +”正全面应用到第三产业，形成诸如互联网金融、互联网交通、互联网医疗、互联网教育等新业态，同时不断向第一产业和第二产业渗透。本章首先结合各国的互联网发展战略，分析和探讨国际互联网行业发展现状，并深入探讨与“互联网 +”密切相关的[大数据](#)、[物联网](#)和[云计算](#)等新一代信息技术；其次对我国“互联网 +”战略进行分析和综述，并针对工业、金融、交通、医疗等典型行业中“互联网 +”模式应用现状进行探讨，进而指出“互联网 +”的发展思路与对策；最后对“互联网 +”的研究进展和发展前景进行展望。

关键词：互联网 +，网络信息服务，互联网金融，互联网交通，大数据

Abstract

With the rapid development and application of Internet, especially the mobile technologiessuch as the Pervasive Computing, Ubiquitous Network, making the Internet increasingly becomes the most important andnecessary production factors for life, economic and social developmentjust as water and electricity. Currently, “Internet +” mode has gradually become a new way of thinkingto re-examine the whole business ecosystem. And itis constantly applied to the tertiary industry, which forms some new industry patterns, such as the Internet Finance, Internet Transportation, Internet Health, Internet Education and others, and in the meantime penetrates the first and secondary industries. This report firstly focuses on the national development strategies of Internet, analyzes the international development status of the Internet industry, and discusses some new information technologies closely related with the “Internet +”, such as Big Data, Internet of Things and Cloud Computing; Secondly, analyzes and reviews the China’s “Internet +” development strategy, and discuss the application status of some typical “Internet +” industries for Finance, Transportation, Health and others; Further points out the “Internet +” development ideas and strategies; Finally, introduces some “Internet +” research progresses and prospects.

Keywords: Internet +, Internet Information Service, Internet Finance, Internet Transportation, Big Date

1 发展背景

随着互联网深入应用，特别是以移动技术为代表的普适计算、泛在网络的发展与

向生产生活、经济社会发展各方面的渗透，物联网、云计算、大数据等新一代信息技术逐渐成为互联网的延伸和发展，在此背景之下，全球数据增长超越了历史上任何一个时期。IDC 在 2011 年的研究报告《从混沌中提取价值》^[1] 中指出，当时全球数据总量为 1.8ZB，预计到 2020 年将增至 35.2ZB，年均增长率超过 40%。其中，信息爆炸式增长最为典型的当属互联网行业，而且这些信息和数据包括不同数据类型，例如：结构化数据、半结构化数据和非结构化数据。据统计，全球每个月发布 10 亿条 Twitter 信息和 300 亿条 Facebook 信息。《福布斯》分析指出，全球 90% 的数据是在过去两年中生成的。现在越来越多新的科学研究领域完全建立在大量数据的基础上，比如系统生物学、宏生态学、基因组学、脑科学等。除此之外，全世界的工业设备、汽车、电表上有着无数的传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，并产生了海量的数据信息。因此看来，大数据已经不同程度地渗透到工业、科技、交通、电力、医疗、金融、社保、国防、公共安全等人类社会的各个行业领域和部门。作为新一轮科技和产业竞争的战略制高点，以大数据、互联网等为代表的信息技术的发展，将推动整个信息产业的创新发展，促进社会生产力的发展，改善人们的生活和工作方式，成为推动世界经济增长和社会发展的重要动力。

在这样的背景之下，互联网产业快速发展，已经渗透到社会生产与生活中，成为信息传播的主要媒介。十二届全国人大三次会议上，李克强总理在政府工作报告中首次提出国家将制订“互联网+”行动计划，加大对互联网行业的管理，同时出台一系列鼓励大众创新、万众创业的举措，对传统行业互联网转型、模式创新提供扶持，推动移动互联网、云计算、大数据、物联网等与现代制造业结合，促进电子商务、工业互联网和互联网金融健康发展，引导互联网企业拓展国际市场。

2015 年 7 月 4 日公布的《国务院关于积极推进“互联网+”行动的指导意见》中指出，“互联网+”是把互联网的创新成果与经济社会各领域深度融合，推动技术进步、效率提升和组织变革，提升实体经济创新力和生产力，形成更广泛的以互联网为基础设施和创新要素的经济社会发展新形态。阿里研究院在《互联网+研究报告》中提到，所谓“互联网+”是指以互联网为主的一整套信息技术在经济、社会生活各部门的扩散与应用过程。此处的“互联网”所指的是以互联网为中心的互联网技术，包括电子信息技术、通信技术、计算机技术、软件技术、网络技术和数字技术、移动互联网、物联网、云计算、大数据、机器认知、语音识别等；而“互联网+”正是以“互联网”技术为基础的创造、创新技术成果与现代经济社会各个领域的大融合、大发展，以推动经济、社会发展为目标，对“互联网”资源全面进行开发与利用。作为一种通用技术，互联网将和 100 年前的电力技术、200 年前的蒸汽机技术一样，对人类经济社会产生巨大、深远而广泛的影响。

2 国际研究现状

2.1 互联网行业发展概述

互联网是近年来继报纸、广播、电视后新生的另一种媒体。随着社会的发展，计算机越来越普及，互联网的功能也越来越丰富。从最初的技术领域应用发展到今天的娱乐休闲应用，网络本身就具有比传统媒体更全面的优势。

在过去 50 年中，每隔 10 年就会因为技术进步产生一个新的计算技术周期。如今，移动互联网将人类带入第 5 个新技术周期的加速发展阶段，其增长速度远超桌面互联网。从 20 世纪 60 年代工业计算起，每一波浪潮大概持续 10 年，从小型机到个人机，从桌面互联网到移动互联网，都符合这个节奏。此外，每一波浪潮都可以支撑 10 倍于前一轮的用户量，因此有着“互联网女皇”之称的华尔街知名分析师 Mary Meeker 也大胆预计，2020 年移动互联网将达到 100 亿台设备的量级^[2]。

近几年，全球互联网行业进入快速发展阶段。在用户方面，互联网用户的渗透率在 20 年间发生了天翻地覆的变化，从 1995 年的 0.6%（3500 万人）发展到 2014 年的 39%（28 亿）。互联网人口构成上也变化显著，1995 年绝大部分是美国（61%）和欧洲（22%）人口，到了 2014 年，亚洲已经占据半壁以上的江山，中国占 23%，亚洲其他国家占 28%，如图 1 所示。

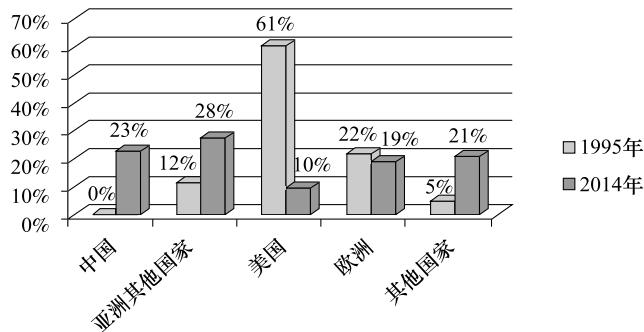


图 1 全球互联网用户比例变化情况

在企业方面，通过市值前 15 位的公司排名即可反映互联网的变迁。1995 年，市值前 15 位的公司里有 13 家美国公司，还见不到一家中国企业的影子。20 年后，美国公司依然强大，市值前 15 位仍占据了 11 席，不过中国公司占据剩下的 4 个席位，分别是阿里 (#3)、腾讯 (#6)、百度 (#8)、京东 (#11)。除了苹果延续强势以外，20 年前的那些公司大部分已经消失或淡出我们的视野，当年的老大 Netscape 只能在教科书中供人铭记，而现在苹果的市值足足是它的 140 倍。尽管互联网已经产生了巨大影响，但是这一切才

刚刚开始。在国防、医疗、教育、政府等领域，以及国民经济第一产业和第二产业中的各个领域，互联网还有很大的发挥空间。

2.2 各国“互联网+”相关战略计划概述与分析

2.2.1 美国：先进制造和工业互联网

金融危机之后，2009年美国提出《重振美国制造业框架》，2011年启动《先进制造业伙伴计划》，2012年正式发布《先进制造业国家战略计划》，这一战略计划在这几年间逐渐清晰。一开始是战略方向，其中提出了12大关键技术，到2013年美国已经聚焦到三大技术的优先突破。这三大技术包括：先进制造的感知控制；可视化、信息化、数字化制造；以及先进材料制造。而这些和德国的工业4.0战略是非常相似的。

承接美国先进制造业的国家战略计划和德国的工业4.0战略计划，GE提出了工业互联网，并于2012年11月发布《工业互联网：突破智慧和机器的界限》白皮书^[3]。工业互联网是全球工业系统与高级计算、分析、感应技术以及互联网连接融合的结果。它通过智能机器间的连接并最终将人机连接，结合软件和大数据分析，重构全球工业，激发生产力，让世界更美好、更快速、更安全、更清洁且更经济。工业互联网将整合两大革命性转变之优势：其一是工业革命，伴随着工业革命，出现了无数台机器、设备、机组和工作站；其二则是更为强大的网络革命，在其影响下，计算、信息与通信系统应运而生并不断发展。工业互联网背后是以工业的互联网为基础，以软件控制应用和软件定义的机器紧密联动，把应用和终端紧密关联起来，然后达到机器之间、机器与人之间，以及企业上下游之间的一种全面的连接交互。最终实现的就是一个以开放和智能为特征的工业体系，这和传统封闭式的工业体系是不同的。GE提出工业互联网以后，得到美国产业界和政府的广泛支持，2013年，美国国家标准研究院把工业互联网纳入其智能制造和CPS系统的专项，开始制定标准的框架，并于2014年4月成立工业互联网联盟（IIC），致力于打破技术孤立壁垒，促进物理世界和数字世界的融合。这个产业联盟的目的是打造一个工业互联网生态系统。这个联盟的核心成员除了GE以外，还有思科、IBM，以网络设备制造商、方案解决商和运营商^[4]。

2.2.2 德国：工业4.0

德国制造业领域技术的渐进性进步被描述为工业革命的四个阶段，即工业4.0的进化历程^②，如图2所示。工业4.0的初步概念是由德国相关协会在2011年德国汉诺威工业博览会首先提出来的。为了在新一轮工业革命中占领先机，在德国工程院、弗劳恩霍夫协会、西门子公司等德国学术界和产业界的建议与大力推动下，德国工业4.0战略在2013年4月的汉诺威工业博览会上被正式推出，并纳入了德国政府《高技术战略2020》

② 来源：德国人工智能研究中心（DFKI），2011。

确定的十大未来项目之中，旨在支持工业领域新一代革命性技术的研发与创新。随后，德国机械及制造商协会等设立了“工业4.0平台”，德国电气电子和信息技术协会也发表了德国首个工业4.0标准化路线图。工业4.0战略有以下几个方面要点：（1）建设一个信息物理系统网络，这是实现工业4.0的基础。（2）研究两个主题，一是“智慧工厂”，重点研究智能化生产系统及过程，以及网络化分布生产设施的实现；二是“智能生产”，主要涉及企业生产物流管理、人机互动以及3D技术在生产过程中的运用。生产流程智能化是实现工业4.0的关键。（3）实现三项集成：通过价值网络实现横向集成，贯穿整个价值链的端到端工程数字化集成，纵向集成和网络化制造系统。总的来看，工业4.0战略的核心就是通过CPS（Cyber-Physics System）网络实现人、设备与产品的实时连通、相互识别和有效交流，从而构建一个高度灵活的个性化和数字化的智能制造模式。

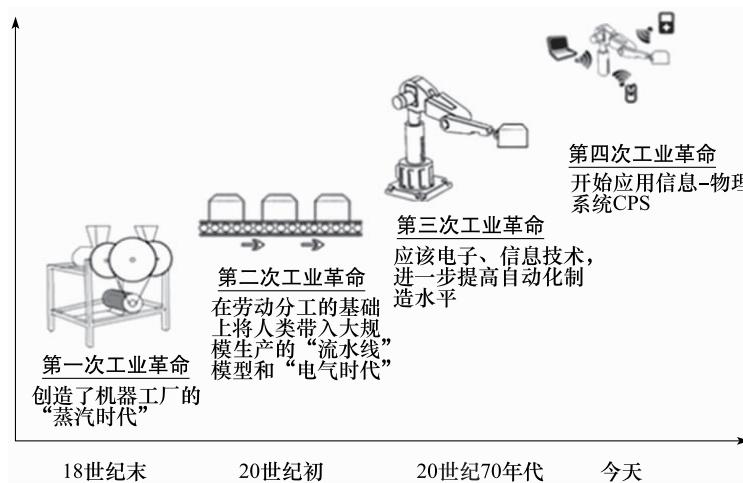


图2 工业革命的四个阶段

工业4.0是结合德国工业的发展阶段，从机械化、电气化、自动化到智能化，实际上也是一个逐步和ICT技术结合的过程。在工业4.0战略中，其中一个核心内容和美国工业互联网非常相似，利用的术语都是CPS，即把资源、信息、机器和人连接起来，构成智能生产与智慧工厂。两者实现的目标也非常相似，即大规模定制以及减少能耗等。德国工业4.0有三大集成，包括工厂内各个环节的集成、企业间的集成，以及整个产品全生命周期的集成。这些集成最后达到的效果就是端到端的数据整合，消除信息不对称，实现整个产品、整个全生命周期的数据连接和传递。工业4.0的核心要素也是三大要素，包括互联网络、工业大数据和智能机器。

2.3 “互联网+”实施关键技术发展现状

2.3.1 大数据

当今全球的数据量已达到ZB级，数据正以前所未有的速度不断增长和累积。其中信

息爆炸式地增长最为典型的当属互联网行业，如图 3 所示。



图 3 互联网中的大数据

学术界、工业界甚至于政府机构都已经开始密切关注大数据问题，并对其产生浓厚的兴趣。英国《Nature》杂志 2008 年“大数据”^[5]专刊集中报道了大数据所带来的技术挑战及未来的发展方向，标志着大数据分析与处理已经成为科学的研究、商业活动、日常生活中一个核心问题，并成为计算机科学研究的最重要内容之一。微软研究院出版的《The Fourth Paradigm》^[6]一书以科学的研究为切入点，阐述了如何在 eScience 时代从事数据密集型的科学的研究。《Science》杂志的 2011 年“数据处理”专刊^[7]主要围绕着科学的研究中大数据的问题展开讨论，阐明大数据对于科学的研究的重要性。2012 年 4 月欧洲信息学与数学研究协会会刊 ERCIM News 上出版专刊“Big Data”^[8]，讨论了大数据时代的数据管理、数据密集型研究的创新技术等问题，并介绍了欧洲科研机构开展的研究活动和取得的创新性进展。自 2012 年以来，大数据的关注度与日俱增。2012 年 1 月的达沃斯世界经济论坛上，大数据是主题之一，还特别针对大数据发布了报告^[9]，探讨了新的数据产生方式下，如何更好地利用数据来产生良好的社会效益。2012 年 3 月美国奥巴马政府发布的“大数据发展计划”^[10]被视为美国政府继信息高速公路计划之后在信息科学领域的又一重大举措。根据该计划，美国国家科学基金会（NSF）、国立卫生研究院（NIH）、国防部（DOD）、能源部（DOE）、国防部高级研究计划局（DARPA）、地质勘探局（USGS）等 6 个联邦部门和机构共同提高收集、存储、保留、管理、分析和共享海量数据所需的核心技术，扩大大数据技术开发和应用所需人才的供给^[11]。过去几年欧盟已对科学数据基础设施投资 1 亿多欧元，并将数据信息化基础设施作为 Horizon 2020 计划的优先领域之一。2012 年专门征集针对大数据的研究项目预算为 5 000 万欧元，仍以数据信息化基础设施为先导。与此同时，联合国的 Global Pulse 倡议项目^[12]阐述了大数据时代

各国特别是发展中国家在面临数据洪流的情况下所遇到的机遇与挑战，同时还对大数据的应用进行了初步解读。

虽然大数据已受到各界广泛关注，但目前对大数据的研究还缺乏科学性、系统性，甚至大数据基本概念、关键技术以及对它的利用上还存在很多疑问和争议。下面内容将阐述大数据分析技术发展现状与趋势。

IBM、Oracle、Microsoft、Google、Amazon、Facebook 等 IT 跨国企业是发展大数据处理技术的主要推动者，部分企业已发布解决方案来应对大数据的挑战^[13]。IBM 将数据分析作为其大数据战略的核心。自 2005 年以来，投资 160 亿美元进行了 30 次数据分析的相关收购，并对其海量数据分析平台 InfoSphereBigInsights 等相关产品进行了一系列创新，以更好地支持大数据处理。Oracle 将数据库作为其大数据战略的中心，将数据挖掘和分析技术整合到现有的数据库产品中，再配合其数据库云服务器、商务智能云服务器以及相关软件，组成大数据系统解决方案。EMC 将云计算作为其大数据战略的平台，推出了基于云基础架构的存储、数据科学协作和自助服务、支持大数据的应用程序等相关产品与服务，使用户从数据源获得最大价值，增强灵活性并提高效率。Facebook 作为社交网络的领导者，积累了海量用户行为和网络群组关系数据，利用用户行为分析，对不同用户群组有针对性地发布广告。Google 针对大数据问题提出了具有代表性的技术——Google 文件系统（GFS）和 Map/Reduce 处理模型，并发表了一系列关于大数据管理和分析处理技术的论文^[14,15]。学术界开展的对于大数据分析的研究尚处于起步阶段。美国太平洋西北国家实验室与华盛顿大学合作成立了大数据研究所，支持大数据科学研究，探索大数据挖掘技术，应对气象变化和能源管理等领域的数据分析。VLDB Journal 2012 年第 21 卷《Large Scale Analytics》专刊探讨了如何利用大规模集群系统所具有的可伸缩性和容错性的优势，实现高效的数据管理功能。大数据技术研究包括 MapReduce^[16,17] 与 Hadoop^[18] 等。

综上所述，大数据技术及相应的基础研究已经成为科技界的研究热点，大数据研究作为一个横跨信息科学、社会科学、网络科学、系统科学、心理学、经济学等诸多领域的新兴交叉方向正在逐步形成。

2.3.2 物联网

物联网的概念来源于美国麻省理工学院（MIT）在 1999 年建立自动识别中心时提出的网络无线射频识别（RFID）系统^[19]——把所有物品通过射频识别等信息传感设备与互联网连接起来，实现智能化识别和管理。早期的物联网是以物流系统为背景提出的，以射频识别技术作为条码识别的替代品，实现对物流系统智能化管理。随着技术和应用的发展，物联网的内涵已发生了较大变化。

2005 年在突尼斯举行的信息社会世界峰会（WSIS）上 ITU 正式确定了“物联网”概念，并随后发布了《ITU Internet reports 2005-the Internet of things》^[20]，介绍了物联网的特征、相关的技术、面临的挑战和未来的市场机遇。ITU 在报告中指出，我们正站在一个新的通信时代的边缘，信息与通信技术的目标已经从满足人与人之间的沟通，发展

到实现人与物、物与物之间的连接，无所不在的物联网通信时代即将来临。

物联网使我们在信息与通信技术的世界中获得一个新的沟通维度，将任何时间、任何地点连接任何人扩展到连接任何物品。万物的连接就形成了物联网。物联网把传统的信息通信网络延伸到更为广泛的物理世界。虽然“物联网”仍然是一个发展中的概念，然而将“物”纳入“网”中，则是信息化发展的一个大趋势。物联网将带来信息产业新一轮的发展浪潮，必将对经济发展和社会生活产生深远影响^[21]。

继美国政府提出制造业复兴战略以来，美国逐步将物联网的发展和重塑美国制造优势计划结合起来以期重新占领制造业制高点。欧盟建立了相对完善的物联网政策体系，积极推动物联网技术研发。德国的工业 4.0 战略利用物联网来提高德国制造业的竞争力，以引领新一轮的工业革命。韩国政府则预见到以物联网为代表的信息技术产业与传统产业融合发展的广阔前景，持续推动融合创新。

受各国战略引领和市场推动，全球物联网应用呈现加速发展态势，物联网所带动的新型信息化与传统领域走向深度融合，物联网对行业和市场所带来的冲击和影响已经受到广泛关注。总体来看，全球物联网应用仍处于发展初期，物联网在行业领域的应用逐步深入，在公共市场的应用开始显现，机器与机器通信、车联网、智能电网等是近几年全球发展较快的重点应用领域^[22]。

2.3.3 云计算

计算机系统的演进过程与现实世界中人类社会的演进过程极其相似，如图 4 所示。人类社会从早期的自给自足的单体生活方式进化到初期部落内协作分工的生活方式，然后演进到城镇化社会，人们共享基本的基础设施与资源，并有着精细化的社会分工协作，最后演进到基础资源高度整合、社会分工高度发达的现代社会；相对应，计算机系统从单机时代进化到能够整合共享资源的专用局域网系统，然后发展到资源可整合、共享的互联网时代，逐步演进到目前资源动态分配、服务高度发达共享的云服务时代。

微软亚太研发集团主席张亚勤认为，云计算是继 20 世纪 80 年代大型计算机到客户端-服务器的大转变之后的又一巨变，被誉为“革命性的计算模型”，它延续了网格计算、分布式计算、并行计算等既有的理论，其远景是以互联网为中心，提供安全、快速、便捷的数据存储和网络计算服务^[23]。权威机构预测，云计算有望成为继大型机、个人机、互联网之后的第四次 IT 产业革命。未来三年中国云计算产业链的产值规模将达到 2 000 亿元^[24]。云计算将 IT 相关的能力以服务的方式提供给用户，允许用户在不了解提供服务的技术，没有相关知识以及设备操作能力的情况下，通过互联网获取需要的服务。云计算具有虚拟化、按需自助、多种网络访问形式、资源共享、弹性快速配置和可度量性等特点。云计算的概念自 2007 年被提出以来，得到了各方的高度重视，并在云平台、云终端、云安全等领域产生了许多有价值的研究成果和商业应用。

从 2007 年开始，云计算在国内进入快速发展阶段。近两年，中国政府对云计算的了解和认可程度不断提高；截止到 2012 年初，北京、上海、成都、杭州、青岛和西安等城市在政府云应用领域进行了探索^[25]。另外，国内主要研究机构、部分省市和大型企业已

已经开始关注云计算产业发展。北京、无锡、广东等省市率先启动云计算基础设施、云计算服务平台和云计算产业园区建设，吸引国内外云计算技术和服务企业入驻，部分云计算平台已开始向企业和社会提供服务。2010年，中关村云计算产业技术联盟成立。中关村云计算产业技术联盟由联想、用友等19家国内单位发起成立。2010年，微软与上海云计算基地合作共建医疗云服务平台。另外，无锡滨湖区搭建“超级云计算”服务平台；青岛软件园搭建“云计算”支撑平台；IBM与苏州国科战略合作，聚焦云计算服务等。2011年10月，国家发改委同财政部、工信部在北京、上海、深圳、杭州、无锡等5个城市云计算服务创新发展试点工作的基础上，启动了涵盖互联网服务、电子商务、金融服务、中小企业服务、公共技术平台等在内的15个重点示范项目。国内很多学术研究机构也都搭建了自己的云计算平台。例如，同济大学嵌入式系统与服务计算教育部重点实验室在2011年创建了包括虚拟超云系统计算平台、存储平台和服务平台在内的同济曙光超云平台。该平台把基于虚拟超市的资源组织与管理思想融入曙光云系统，以保证计算平台获取资源的优先性。该平台主要为学校智能交通、远程医疗健康、生物信息等领域的科研提供云计算服务。东南大学参与到诺贝尔物理学奖获得者丁肇中教授领导的国际合作项目AMS-02(Alpha Magnetic Spectrometer 02)试验中。为了满足AMS-02海量数据处理应用的需求，东南大学构建了相应的云计算平台，平台分别从IaaS、PaaS和SaaS三个层次提供海量数据处理服务。

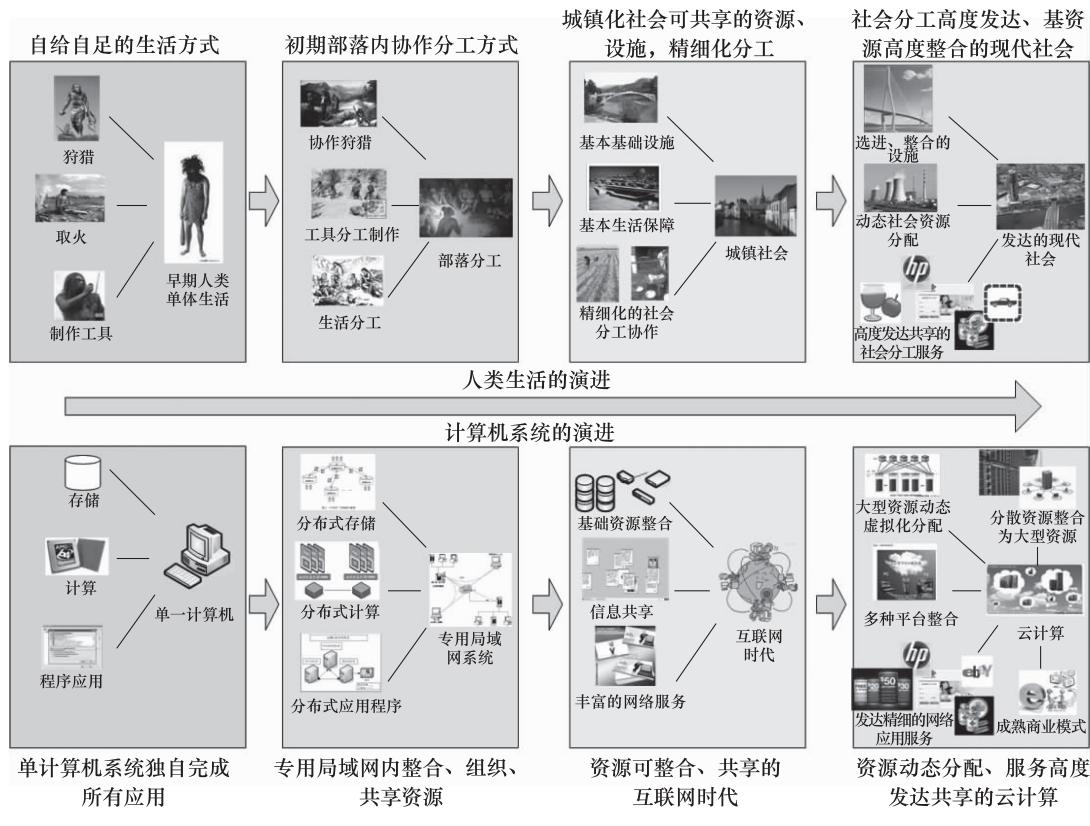


图4 计算机系统的演进过程

国内的互联网巨头如百度、腾讯、阿里巴巴等也相继打造了各类云服务和云应用平台。例如，百度应用引擎（Baidu App Engine）提供了 PHP、Java、Python 的执行环境，以及云存储、消息服务、云数据库等全面的云服务^[26]。腾讯云计算平台提供了第三方应用从 Web 接入到上线运营整个过程中涉及的一系列服务，以降低技术和运营门槛，节省开发者的运营成本，提升运营效率，可为广大第三方应用开发和运营团队创造更多价值^[27]。阿里巴巴打造的阿里云为用户提供了弹性计算、云存储、云引擎等其他一系列的云服务^[28]。另外，软件厂商如用友、金蝶、金山等都推出了自己的 SaaS 应用。中国万网、世纪互联、网宿科技、蓝汛、盛大云等都相继推出 IaaS 服务。其中，万网首先引入了“虚拟化技术”，将旗下几乎所有应用产品全部转型为云服务^[29]。万网云计算基于多台大型服务器集群构架而成，采用 BGP 骨干网络，异地机房之间采用千兆光纤连接，实行企业级虚拟化技术和管理平台，实现集群内的弹性可伸缩配置。

智能移动通信云终端市场也成为国内各大互联网厂商抢占的对象。小米科技公司用小米手机搭载 MIUI 系统和米聊软件，以较高的性价比迅速吸引了一批用户，并赢得了智能手机市场一定的份额。阿里巴巴与国内手机厂商天语合作，推出了阿里云手机，该手机内置了几乎所有的淘宝应用。奇虎 360 公司联合 TCL、海尔等公司，搭载其云安全软件，面向学生群体推出 360 智能手机。百度联合长虹研发千元智能手机，用户可以方便简易地使用百度云平台和服务。盛大联合爱立信联合开发 Bambook 智能手机，其中包含云中书城、盛大网盘等盛大云服务终端产品。

2.3.4 实时并发

实时系统在生产生活的各个领域都有重要应用。而在当前“互联网+”时代，随着数据量的激增及应用种类和形态的多元化，系统的实时性正面临更大的挑战，要求系统具有更高的并发处理能力、实时风险防御能力和资源动态配置能力等。例如，在互联网金融领域，只有通过实时的风险辨识才能保证网络交易用户使用体验的同时确保交易安全，而在智能交通领域中，同样需要交通道路实时路况的计算和实时控制。在这些应用中，如果系统在时间限内不能及时完成任务，就会导致系统瘫痪甚至是灾难^[30]。

实时系统的另一特点是通常以多任务的并发方式进行。在并发系统中，如果每个任务都在不同的处理器上同时执行，那么并发就是真实的。相反，如果各个任务之间是交错运行的，只是在同一处理器的不同时间片上执行，那么这种并发就是虚拟的。实时系统必须保证各个任务的每一次请求都在时间限内完成，因此任务调度问题是实时系统研究的重点。研究者们分别设计了多种可行且高效的任务调度算法，在工业生产、公共交通、网络路由及其他高科技领域都有重要的应用^[31]。

由于调度问题本身是 NP 完全问题，国内外的研究者提出了很多启发式算法。根据算法基本思想的不同，传统的并行静态任务调度的算法大致可分为 4 类：表调度算法^[32-34]、基于任务复制的调度算法^[35,36]、基于任务聚类的调度算法^[37,38]和基于随机搜索的调度算法^[39]。根据对目标系统的假设不同，静态任务调度又可分为同构环境下的任务调度和异构环境下的任务调度两种。表调度的基本思想是：通过对节点分配优先级别来构造一个

调度列表，然后重复从调度列表中顺序取出第一个节点，将节点分配到使它的启动时间最早的处理器上。直到任务图中所有节点被调度完毕。大多数表调度算法假设目标系统是处理单元数目有限且完全连接的同构环境。通常，表调度算法比较实用，与其他调度算法相比，其时间复杂度相对较低，调度结果较好。基于任务复制的调度的基本思想是，在一些处理器上冗余地映射任务图中的一些任务，以达到减少处理器之间通信开销的目的，即它利用处理器的空闲时间复制前驱任务，可以避免某些前驱任务的通信数据的传输，从而减少处理器等待的时间间隙。其目标系统一般是处理单元数目不受限制的同构环境。基于任务聚类调度的基本思想是把给定任务图的所有任务映射到数量不受限的集群上。如果两个任务分配到同一个聚类，则表示它们在同一个处理器上执行。除了执行聚类的步骤外，算法还必须对完成映射的聚类进行最后的调度，即对聚类中的任务在每个处理器上根据时间先后进行排序。基于随机搜索技术的调度算法主要通过有导向的随机选择来搜索问题的解空间，而并不是单纯的随机搜索。这类技术组合前面搜索结果的知识和特定的随机搜索特点来产生新的结果。其主要代表是遗传算法和模拟退火方法。遗传算法比一般的启发式算法的调度结果要好，然而，它们往往有较高的时间复杂度，而且需要适当地确定一些控制参数。此外，适用于一个任务图的最优控制参数往往不适合另一任务图，很难找到对于大多数任务图都能产生较好的调度结果的控制参数集。当前，异构环境下任务调度的研究往往是针对元任务或者批任务开展，忽视了任务之间的数据关联和优先约束关系，即大多数针对独立任务的调度不能反映异构环境下任务的实际特征。早期的大多数并行任务调度算法往往基于较为简单的假设，例如假设任务执行时间相同、任务之间无通信、处理器完全连接以及处理器数目不受限制等，而且，它们通常缺乏对异构资源特征的分析。文献[40, 41]利用模糊聚类的理论分析异构资源特征，并根据异构资源特征分析的结果提出启发式调度算法。

2.3.5 应变适配

在当今“互联网+”环境下，“触网”的物体和对象日益增加，这使得存在于互联网络上的信息服务的内涵越来越丰富，互联网络日趋庞大和复杂，参与的用户也越来越多，随之而来的是用户需求的多样性和互联网络环境的复杂多变性。在此趋势下，网络信息服务的应变与适配技术就显得越来越重要。近年来流行的P2P服务、网格服务、面向服务的构架(SOA)和云计算等信息服务技术为网络环境下跨域、跨组织的应用集成和信息服务带来了巨大的机遇。P2P服务是一种面向互联网的应用，参与对等计算的各个结点作为平等的对等体，通过直接交换来共享资源和信息^[42]。网格服务以20世纪90年代后期开始的网格计算为代表，网格服务的目标是共享和整合广泛分布的网络资源，为用户提供虚拟网络计算环境^[43]。随着电子商务的迅速崛起，Web服务作为一种新兴的Web应用模式应运而生^[44]。Web服务是Web上数据和信息集成的有效机制，它基于一系列开放的标准技术，具有松散耦合、语言中立、平台无关性和开放性。Web服务和相关技术的不断发展促使SOA逐渐盛行。SOA平台下，用户、过程、应用和数据被全面整

合起来。服务将分布、异构的资源整合起来，呈现为统一的逻辑对象，在开放的标准之上，以安全和可管理的方式供用户使用。近年来，分布式处理、并行处理、网格计算等技术的发展促使云计算这一计算模式和应用平台的产生。在云计算模式下，功能强大的应用和服务被整合打包发布到互联网上，以便用户直接使用，这一方式具有巨大的发展潜力^[45]。同时，这些信息服务技术也分别从不同侧面来努力实现应变和适配：网格服务通过异构集成来实现资源的共享和灵活应用，SOA 和 Web 服务通过服务的重用和组合来实现服务系统灵活架构和柔性服务流程，云计算通过虚拟化等技术来解决系统和资源的可扩展性与伸缩性等。

综上所述，现有网络信息服务技术侧重强调和解决资源、服务的互联互通，通过资源聚合和协同的方式实现上层应用，有力地推动了网络信息服务的发展，并且针对应变与适配的需求，它们也进行了各自的努力和尝试。但是，在当前“互联网+”环境下，需求的多样性、环境的多变性内涵更加丰富，对信息服务技术也提出了更高的要求。需求的多样性要求信息服务技术能够按需聚合相应的服务资源，并保证服务流程的正确性和所提供的内容的精准性。环境的多变性要求信息服务能适应异构环境下软件、硬件资源的变化特征，做到与环境的适配性。针对此挑战，以同济大学为牵头单位的973项目《信息服务的模型与机理研究》从网络环境下信息服务的“流程、内容和环境”三要素及其相互关系出发，凝练出其中的两个关键科学问题：网络信息服务的表达性问题和适配性问题，如图5所示。表达性问题是针对不确定和多样的用户需求准确地设计和表达服务的流程与内容，以提供动态、精准、可伸缩的信息服务，满足用户需求的问题。适配性问题是针对异构、复杂、动态的网络环境实现和增强服务对环境的适配能力，促进服务聚合和协同，提升信息服务质量问题。围绕关键科学问题，以“变和应变”为思路，构建网络信息服务的过程范式基础理论，提供服务实施过程中的一系列如流程设计、环境协调、质量保证等关键技术，以及网络信息服务的运行支撑平台，实现与增强互联网环境下信息服务的应变与适配能力^[46]。

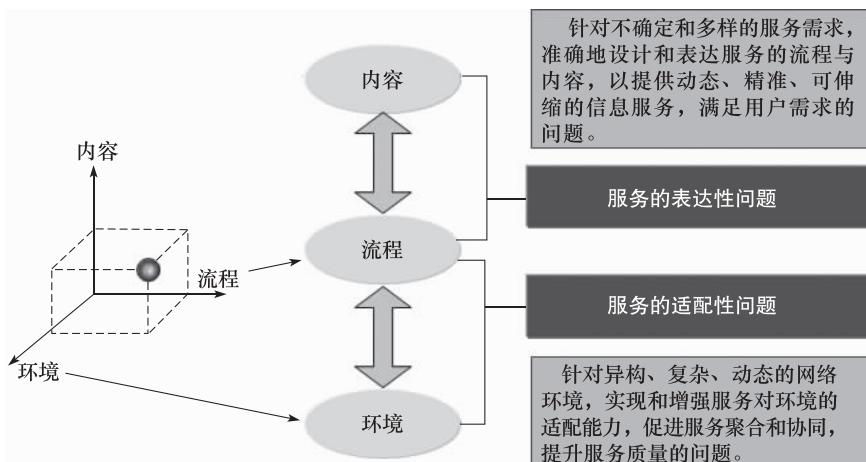


图5 973项目“信息服务的模型与机理研究”关键科学问题

3 国内研究现状

3.1 我国互联网行业发展概述

1994 年我国通过一条 64K 的国际专线全功能接入国际互联网，中国互联网时代从此开启。近年来，中国网民规模不断增大。根据 CNNIC 发布的《第 35 次中国互联网络发展状况统计报告》数据显示^[47]：2014 年中国网民规模达到 6.49 亿人，手机网民为 5.57 亿人，互联网的普及率达到 47.9%，比 2013 年底提升 2.1 个百分点，2015 年将达到 7.31 亿人。艾瑞咨询统计的数据显示，2014 年中国移动互联网市场规模达到 2 134.8 亿元，同比增长 115.5%，同时未来依旧会保持高速增长，预计到 2018 年移动互联网市场规模将突破 1 万亿大关。

在过去的 20 多年的时间里，互联网成为中国经济全球崛起最好的催化剂。第一个 10 年里，互联网和传统行业和平共处，催生了很多新经济形式，比如门户、游戏和电商等；第二个 10 年里，互联网逐步开始改变甚至颠覆了很多传统行业，分别经历了 4 个阶段：最初，人们使用互联网进行营销，在互联网上发布广告；紧接着，网购、电子商务快速兴起；进而，厂商开始按照互联网的特点生产产品，比如小米手机；到当前阶段，互联网已经开始全面改变和颠覆各个传统行业，包括金融、教育、旅游、健康、物流等。

随着我国互联网基础设施不断完善和提升，应用创新和商业模式创新层出不穷，互联网企业掀起新一波上市浪潮，消费互联网迅猛增长，产业互联网化步伐进一步加快，互联网向金融、交通、教育、影视等传统领域加速渗透，互联网领域管理统筹协调能力大幅增强。展望未来，我国互联网将进入一个崭新的发展阶段，网络基础设施建设将更加集约共享，应用创新和商业模式创新将继续层出不穷，传统产业将全面拥抱互联网，跨界融合范围将持续扩大，互联网金融在政策上有望取得突破，互联网和物流企业将加大对服务资源的整合力度，众筹、众包等模式将创造出更多的行业新业态。

3.2 “互联网+”战略

3.2.1 “互联网+”

随着互联网深入应用，特别是物联网、云计算、大数据等新一代信息技术作为互联网的延伸和发展，同时向人类生产生活、经济社会发展各方面渗透，互联网像水电一样逐渐成为各行各业所必需的生产要素。互联网思维已经逐步被越来越多的企业家，甚至企业以外的各行各业、各个领域的人所认可，成为一种对市场、用户、产品、企业价值

链乃至整个商业生态进行重新审视的思考方式。当前互联网思维正全面应用到第三产业，形成诸如互联网金融、互联网交通、互联网医疗、互联网教育等新业态。

从全球范围来看，互联网领域过去 15 年中已经孕育了大量大市值公司，如美国的苹果、谷歌、亚马逊、微软、甲骨文等形成了新的千亿美元市值企业。而伴随经济持续高速增长，中国互联网行业发展也进入全面化、多元化、国际化阶段。目前全球市值前 10 名的互联网企业中，中国占据 4 席；在前 20 名中，中国有 6 席。这充分说明，随着传统行业与互联网的跨界、融合和创新热潮，中国互联网企业的飞速发展，正在让中国成为互联网领域的中坚力量。

虽然互联网和移动互联网已经给我们的经济和社会带来了巨大的改变与深远的影响，经过多年创新发展的中国互联网企业已跻身世界前列，为我国信息经济发展奠定了坚实基础。但是国内很多行业由于缺乏对“互联网+”的正确认知，认知的不足阻碍了某些行业及相关企业借助“互联网+”的手段来提升生产效率，实现转型升级，激发新的活力。所以，应当深入贯彻国家的“互联网+”发展战略，推动“互联网+”健康发展，深入促进互联网与各产业融合创新，在技术、标准、政策等多个方面实现互联网与传统行业充分对接。

3.2.2 中国制造 2025

2013 年中国工程院会同工信部等启动了“制造强国战略研究”重大咨询项目，由中国工程院院长周济亲自挂帅，组织 50 多位院士和 100 多位专家开展调研，提出在 2025 年进入制造强国行列的指导方针和优先行动。此后，工信部、国家发展和改革委员会、科技部和国资委联合编制的《中国制造 2025》规划，经李克强总理签批，已由国务院于 2015 年 5 月 8 日正式公布。

如果说德国的工业 4.0 是德国在面对美国的信息产业崛起和中国的制造成本两方面的侵袭下，试图摸索未来工业生产的途径、重建产业优势的战略选择，那么“中国制造 2025”规划则是中国版的“工业 4.0”规划，提出了中国制造强国建设三个 10 年的“三步走”战略，为把我国打造成现代化的工业强国描绘出清晰的路线图，代表了中国在由制造大国向制造强国转型过程中的顶层设计和路径选择，意义深远重大。

早在 2002 年，中国就提出了“两化融合”的概念，之后“以信息化带动工业化、以工业化促进信息化，走新型工业化道路”成为国内工业、制造业开展信息化和工业化融合的主要制造方针，有力推动了工业化和现代化进程，显著增强了综合国力。同时，随着大数据、物联网、智能终端、工业互联、移动宽带在制造业的应用，传统的生产及商业模式发生了巨大的变化。“中国制造 2025”是“互联网+”重要的一部分，互联网技术降低了产销之间的信息不对称，加速两者之间的相互联系和反馈，因此，催生出消费者驱动的商业模式。“中国制造 2025”代表“互联网+制造业”的智能生产，孕育大量的新型商业模式，只有将信息化、网络化、智能化与工业生产相融合，才能提升效率、加强竞争力，实现“中国制造”向“中国智造”转型。当前，新一轮科技革命和产业变革与加快转变经济发展方式形成历史性交汇，这也是中国制造业创新驱动、转型升级的发展方向。

3.3 典型行业“互联网+”应用现状

当前随着我国互联网基础设施的不断完善和提升，应用创新和商业模式创新层出不穷，互联网企业掀起新一波上市浪潮，消费互联网增长迅猛，产业互联网步伐进一步加快，互联网加速向金融、交通、教育、影视等传统领域渗透，互联网领域管理统筹协调能力大幅增强。2015年7月4日公布的《国务院关于积极推进“互联网+”行动的指导意见》明确了未来3年以及10年的“互联网+”发展目标，提出包括益民服务、便捷交通、普惠金融等11项重点行动，这些行动涵盖制造业、金融等具体产业，也涉及医疗、教育、交通等民生方面。其中“互联网+”协同制造行动中，在重点领域推进智能制造、大规模个性化定制、网络化协同制造和服务型制造，打造一批网络化协同制造公共服务平台，加快形成制造业网络化产业生态体系。工业互联网正在从消费品工业向装备制造和能源、新材料等工业领域渗透，全面推动传统工业生产方式转变；农业互联网也从电子商务等网络销售环节向生产领域渗透，为农业带来新的机遇，提供广阔发展空间。

目前在民生、医疗、教育、交通、金融等领域，互联网对传统行业的提升作用越来越明显。在民生领域，截至2014年底，各级政府已经在微信上开通了近2万个公众账号面向社会提供各类服务。广州、武汉、上海、杭州等城市已成为微信“智慧城市”，已经有91万广州市民通过微信上的“城市服务”入口获得医疗、交管、交通、公安户政、出入境、缴费、教育、公积金等17项民生服务。在教育领域，阿里云近日宣布启动阿里云大学合作计划，联合8大高校开设云计算与数据科学专业，预计在3年内完成100所高校的专业课程开设，拟培养5万名云计算和数据科学工作者。在医疗领域，全国已有超过100家医院上线微信全流程就诊，超过1200家医院支持微信挂号，服务累计超过300万患者，为患者节省超过600万小时，大大提升了就医效率，节约了公共资源。在交通领域，快的、滴滴打车等移动互联网应用，正在改变人们的出行方式；携程、去哪儿、12306等已经实现了航空、铁路的网上便捷购票；百度地图、高德地图成为出行者导航的首选。在金融领域，随着传统的金融机构与互联网企业的融合，不断推出创新的金融模式，如网络借贷平台、众筹模式的网络投资平台、手机理财APP以及第三方支付平台等。未来互联网金融将在促进经济增长的同时，大幅减少交易成本，提供一站式服务。

4 “互联网+”战略发展思路与对策

4.1 数据资源层面的“互联网+”

大数据代表了“互联网+”的数据资源层，是互联网智慧和意识产生的基础。目前

大数据研究主要侧重于数据的存储、加工与分析，这些研究推动了数据存储技术和数据分析技术的发展和创新。但是，随着互联网上大数据爆炸式增长，多样化趋势等特征越来越显著，互联网的庞大容量和广泛用户基础使得这些大数据的广度和全面性远远超越了一个或几个公司和机构能够人工收集的程度，也使得在这些数据上进行的分析结果尤为深刻和重要。现有的方法由于缺少对互联网数据整体上的考虑，无法刻画和度量互联网上数据资源的总体分布和数据成分等特征。因此，基于互联网大数据的资源特征，大数据分析的首要任务是通过数据“勘探”的方法，对大数据“矿产”形成一个宏观上的认识，通过抽样查询方式获得有效样本，并在样本数据表示与度量的基础上得到数据资源的准确估计，实现能效优化的在线勘探分析。

4.2 系统层面的“互联网+”

信息与网络技术的发展正处于一个极其迅速的重要变革期，其一，终端技术、网络技术、业务平台技术呈现异构化、多样化的趋势。其二，数据的复杂性增加且更加多样化，如文本、图像、语音与视频等；海量数据；人机交互更加频繁与密切。其三，基于网络的信息服务规模尺度和复杂程度与日俱增，网上大量用户不确定的多样性服务要求导致频繁变化与多目标的系统需求。上述发展趋势给“互联网+”提出巨大的挑战，即如何实现整体认知、快速辨识、实时优化是目前“互联网+”面临的重大挑战。这就要求我们结合系统论和还原论，构建多维、多粒度的系统组织网络，实现系统整体的关联性认知和理解，揭示系统所具有的一般性规律，进而提供新型适用于“互联网+”的计算模型和反馈控制技术，确保系统实时精准，满足系统需求。

4.3 面向全产业链的“互联网+”

当前在顺应“互联网+”的潮流中，发展比较快速的行业主要集中在第三产业服务业，但从一些国家的战略来看，如德国的工业4.0、美国的工业互联网以及中国制造2025重点聚焦于第二产业工业尤其是制造业的“互联网+”。例如纺织行业，涉及从棉花种植、原材料供应、纺纱织布、染整、成批生产，到品牌营销管理、渠道管理、物流配送、零售终端等诸多环节，即从农业到工业再到服务业贯穿了第一、二、三产业。要实现面向全产业链的“互联网+”战略，如图6所示，应全面考虑所涉及的三大产业，打通全链条的各个环节，并进行商业模式上的创新，从过去劳动密集型为主的产业向资本密集型产业转移，特别是从传统的产业向战略新兴产业转移。同时，借鉴苹果App Store、谷歌Android市场模式，充分发挥群体大众智慧的力量，形成关于用户各种创意及需求的收集源，集聚融合各种新模式、新机制、新服务、新文化等创新资源，面向整个产业链上下游，打造行业创新创业平台，为该行业全链条上的万众创业与创新提供平台，最大限度地发挥资源的作用，提高资源利用率，形成开放式创业生态系统。



图 6 面向全产业链的“互联网+”战略

5 “互联网+”研究进展与展望

在“互联网+”时代，数据已经渗透至各个行业，并且呈现出数量大、动态性、类型复杂等显著大数据特征。目前已有大数据分析平台的研究工作侧重于大数据管理、处理、分析和可视化中的一个或两个。但是，随着大数据爆炸式增长，多样化趋势等特征越来越显著，现有的方法本质上缺少对数据整体上的考虑，无法刻画和度量数据资源的总体分布和数据成分等特征。基于这样的考虑，大数据分析的首要任务是通过数据“勘探”的方法，形成大数据资源宏观上的认识^[47,48]。为此提出了一个基于索引网络的大型数据资源服务框架，其中包括三个主要部分：数据资源识别与获取、数据资源存储与分析、构建服务支撑平台，如图 7 所示。

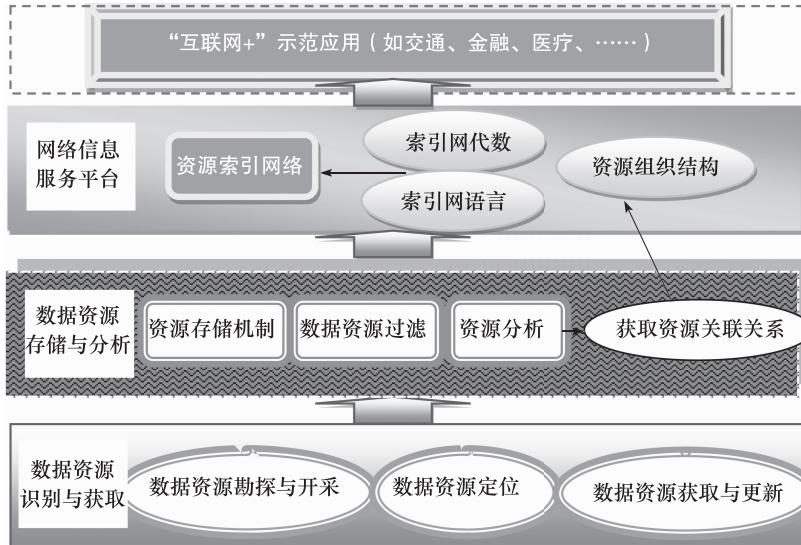


图 7 大型数据资源服务架构

- 数据资源识别和获取。大型数据资源通常是分散的、异构的，而且由于数据量非常大，数据完全获取的方式显然是不可能的，需要通过抽样的方法，获取少量有效样本以统计出总体分布。因此，在数据资源识别和获取这一层次，一方面，将通过探讨所访问的互联网资源的类型、数据成分、网络接口限制等特点，正确分析这些因素对于数据获取和分析的影响，建立符合大规模网络数据资源特性的统计模型；另一方面，将在综合考虑各种网络限制的基础上，通过数据资源勘探和探索等方法，引入拒绝抽样等技术以确保样本单元的独立性。
- 数据资源存储和分析。目前海量异构数据一般采用分布式存储技术，如 GFS 和 HDFS，但它们仍不能解决数据的爆炸性增长带来的存储问题，静态的存储方案并不能满足数据的动态演化所带来的挑战。因此，在数据资源存储和分析这一层次，需要根据特定的数据资源建立相应的分析和存储方法。一个良好的存储机制可以从多样化的方面支持资源分析。而资源分析的目的是提取数据资源之间的关联。其中，针对复杂数据分析方法有助于从多个数据源推断出聚集的分析结果；而针对非结构化和半结构化数据，为了得到更有价值的数据信息分析结果，需要借助于机器学习等语义分析技术，获取数据资源之间的语义和逻辑关系。
- 网络信息服务平台。当前，网页是互联网服务中最基本的资源，在信息呈现、支持应用程序和提供服务等方面发挥主导作用。每天都有众多的网页加入互联网中，其中大部分是冗余的、无序的。因而从互联网上查找所需的服务资源是非常有挑战性的。基于在 Web 超链分析领域的深入研究，以及将网页之间的超链接视作现实世界的客观关系，建立基于网页的分类和超链接分析的索引网络模型^[49]，并给出其代数运算的定义。索引网络支持根据具体要求获取服务资源，以及寻找它们之间的语义关联，能产生更丰富的知识和有价值的信息服务^[50]。文献 [51] 对这一原型系统进行了更深入探讨。

5.1 互联网交通

在交通领域，发展“互联网+”战略，借助移动互联网、云计算、大数据、物联网等先进技术和理念，将互联网产业与传统交通运输业进行有效渗透与融合，形成新业态和新模式，满足公众更便捷出行、更人性服务和行业更科学决策的需求，加快推进交通运输由传统产业向现代服务业转型升级。

前期工作从应用支撑的角度出发，以智能交通领域为应用背景，验证流融合机的模型与基础理论，实践软硬网络的协同虚拟与海微流的协同处理等关键技术，构建面向交通监管的城市智能交通协同监管与实时服务平台，研制面向交通安全的移动车辆同步跟踪实证系统。

城市智能交通协同监管与实时服务平台体系结构如图 8 所示，共分为以下 5 个层次：

- 基础网络层。智能交通领域涉及繁杂多源的网络接入设备，包括摄像头、路口信号机、各种传感器、车检器、移动终端等，涵盖的网络包括交通/公安专网、以

太网、无线传感器网络、GSM/CDMA 移动通信网络等分属计算网络、通信网络和控制网络的异构网络环境。基础网络层通过软硬网络的协同虚拟技术，构建统一的面向城市智能交通协同监管与实时服务的网络环境，以满足移动目标同步跟踪的实时性、动态性和分布性等特点，有效集成跨网络、跨地域的视频监控。

- 技术支撑层。开发一组面向交通监管的海量视频信息分析处理关键技术与子系统，包括视频采集、视频管理、视频分析处理等，为融合网络环境下视频监管的开发、部署和运行提供支持。
- 应用支撑层。开发一组面向相交通监管的海量视频信息提取方法与相应子系统模块，包括车辆目标检测、队列长度检测、车牌识别、流量检测、事件检测等。
- 核心应用层。平台的核心应用主要包括两个方面，一是根据视频检测的流量或车队长度信息，实现基于模糊推理的自适应交通控制；二是引入具有计算能力的智能视频采集设备，分布的智能视频采集设备主动构建自组织传感网络，实现实时感知异常事件的发生、在线分析异常事件特征，并主动实时反馈给自组织网络中的其他智能视频采集设备以便联合接力跟踪，最终实现面向交通安全的移动车辆同步跟踪。
- 平台表现层。平台展示给用户的主要功能包括路况实时在线监测、信号灯控制、肇事事件自动检测与报警、肇事车辆轨迹回放等。

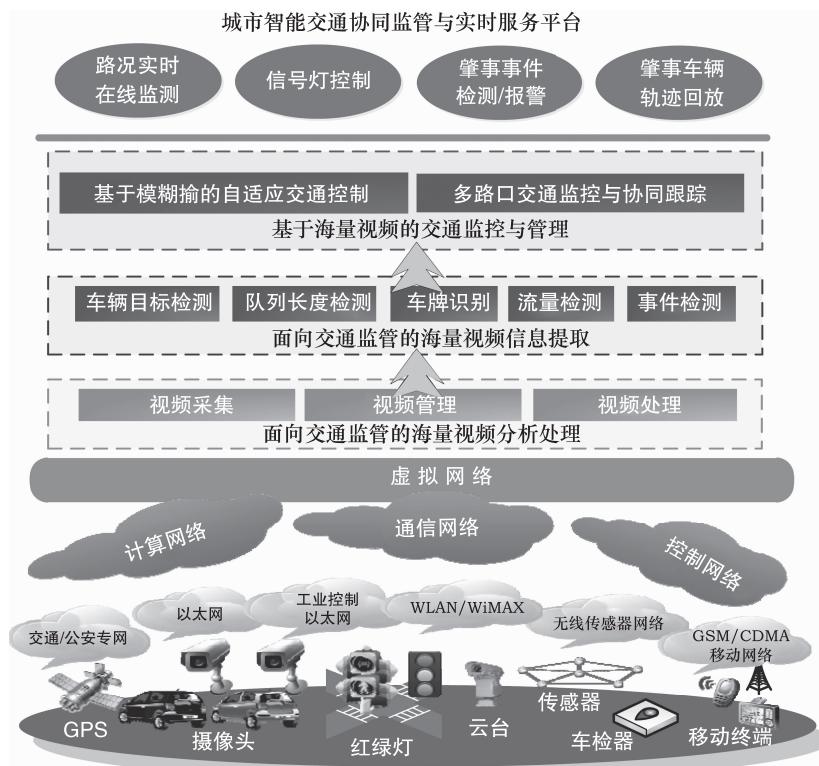


图 8 平台体系结构

5.2 互联网金融

在金融领域，发展“互联网+”战略，鼓励互联网与银行、证券、保险、基金的融合创新，为大众提供丰富、安全、便捷的金融产品和服务，更好满足不同层次实体经济的投融资需求，培育具有行业影响力 的互联网金融创新型企业。

在互联网金融环境中，数据作为金融核心资产，具有相当大的价值，但同时它又存在着巨大的安全隐患。金融行业不能容忍任何安全问题，一旦出现问题，必然会对企业和个人造成巨大的损失^[52]。针对网络金融信息安全问题，研究并开发以行为认证为核心的可信网络金融交易系统，围绕软件行为认证等关键技术，搭建行为认证平台体系。

在认证中心搭建过程中，通过在用户安全客户端以及电商网站和支付平台部署行为监控器，形成网络交易可信认证系统平台，并制定网络交易可信认证的认证协议。在网络交易可信认证系统中，认证中心主要负责管理用户行为和软件行为证书，同时实时认证软件及用户行为的可信性^[53]。

网络交易可信认证中心^[54]底层支持多种操作系统，具有良好的跨平台能力。系统之上的支撑技术为上层的应用开发提供良好支持。在支撑技术之上设计通信管理模块、证书管理模块和数据库管理模块；通信管理模块能够针对该系统特定需求对网络通信功能进行封装，为上层提供数据交换等通信服务；证书管理模块对软件行为证书、用户行为证书以及数字证书进行统一管理，包括证书的搜索、更新、发布等操作；数据库管理模块负责更新和维护数据库，提高数据访问效率。基础管理模块之上就是网络交易可信认证系统的第四方认证域，其主要功能是监控和认证网络交易过程，对交易三方进行数字认证、通过用户行为证书验证用户身份的可信性、通过软件行为证书验证交易三方的网络交易行为的可信性。网络交易可信认证中心架构如图9所示。

可信认证中心监控中心属于可信网络交易软件系统试验环境与示范应用项目，用于监控用户、商家和第三方支付公司在进行在线交易行为时产生的用户行为数据与软件行为数据，并采用多种类多维度的表格与图表的方式直观动态展现交易过程中产生的数据。监控中心作为直观动态展现以上数据的平台，目前主要分为三个部分，每个部分又分别由三个屏幕组成，总共由9个屏幕组成。三个部分分别为平台软件行为监控、平台交易数据监控和平台用户行为监控。软件行为监控分屏显示包括购物者、电商、第三方支付平台三方的软件行为监控日志。平台交易数据监控为模拟经过第四方认证平台的实时交易模拟数据，具体包含了滚动展现的交易日志、全国交易数据的分布以及平台实时的交易额与交易笔数数据。用户行为监控以单用户与多用户的用户行为浏览日志与评分，以及包含频繁访问类和访问时间段在内的多维度的用户浏览习惯。

第一部分为平台软件行为监控，其主要目的是监控电商、第三方支付以及用户的软件行为。监控系统通过滚动列表的方式展示软件行为的日志，并高亮显示异常交易，以此帮助业务人员分析异常报警。部分界面如图10所示。

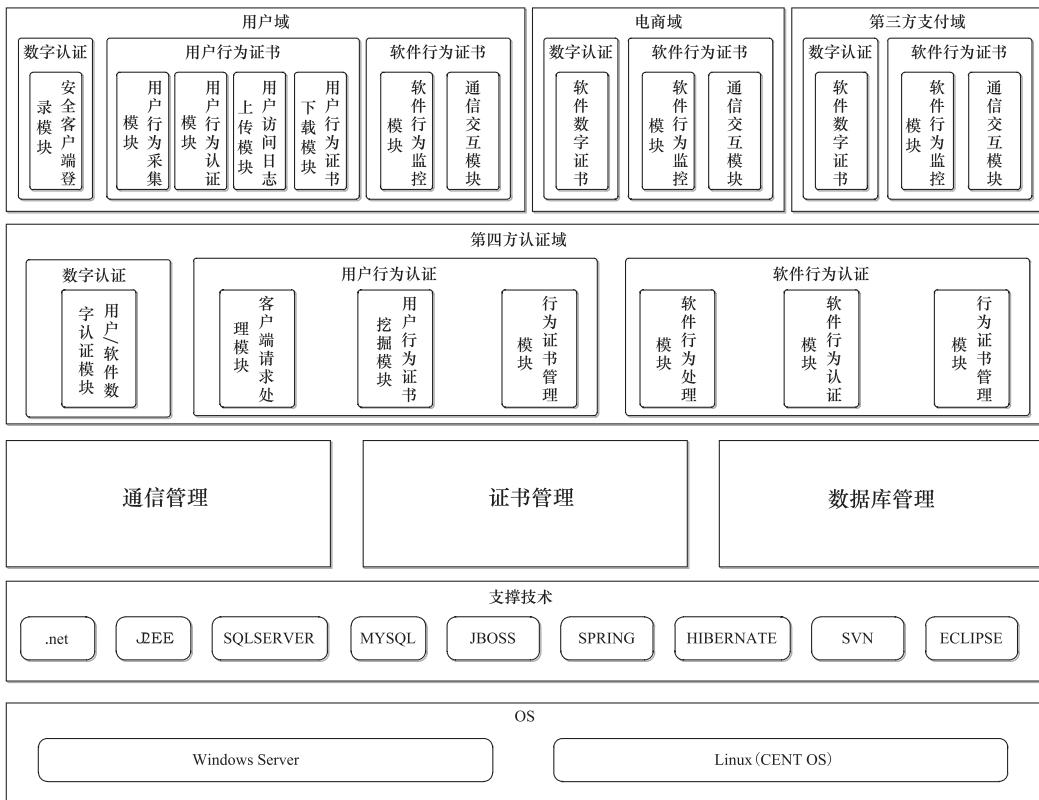


图 9 网络交易可信认证中心架构图

商家软件行为监控							
ID	订单ID	发生时间	输入标识	输入URL	输出标识	输出URL	发生变迁
4237	520	2013-09-04 13:37:18 Wed	1	orderIdgrossstatusid:	0	null	HandleIPN 异常
4234	520	2013-09-04 13:37:09 Wed	0	null	0	null	PlaceOrder 正常
4232	519	2013-09-04 13:36:58 Wed	1	10.60.149.195/junzhu_	0	null	FinishOrder 正常
4229	519	2013-09-04 13:36:50 Wed	1	orderIdgrossstatusid:	1	http://10.60.149.179/	HandleIPN 正常
4226	519	2013-09-04 13:36:41 Wed	0	null	0	null	PlaceOrder 正常
4224	518	2013-09-04 13:36:30 Wed	1	orderIdgrossstatusid:	0	null	HandleIPN 异常
4221	518	2013-09-04 13:36:21 Wed	0	null	0	null	PlaceOrder 正常
4219	517	2013-09-04 13:36:10 Wed	1	10.60.149.195/junzhu_	0	null	FinishOrder 正常
4216	517	2013-09-04 13:36:01 Wed	1	orderIdgrossstatusid:	1	http://10.60.149.179/	HandleIPN 正常
4213	517	2013-09-04 13:35:52 Wed	0	null	0	null	PlaceOrder 正常
4211	516	2013-09-04 13:35:42 Wed	1	10.60.149.195/junzhu_	0	null	FinishOrder 正常
4208	516	2013-09-04 13:35:33 Wed	1	orderIdgrossstatusid:	1	http://10.60.149.179/	HandleIPN 正常
4205	516	2013-09-04 13:35:21 Wed	0	null	0	null	PlaceOrder 正常

第三方支付款软件行为监控							
ID	订单ID	发生时间	输入标识	输入URL	输出标识	输出URL	发生变迁
4230	519	2013-09-04 13:36:53 Wed	1	http://10.60.149.179/	0	null	Paying2 正常
4228	519	2013-09-04 13:36:47 Wed	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4217	517	2013-09-04 13:36:04 Wed	1	http://10.60.149.179/	0	null	Paying2 正常
4215	517	2013-09-04 13:35:58 Wed	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4209	516	2013-09-04 13:35:36 Wed	1	http://10.60.149.179/	0	null	Paying2 正常
4207	516	2013-09-04 13:35:30 Wed	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4201	515	2013-09-04 13:35:07 Wed	1	http://10.60.149.179/	0	null	Paying2 正常
4199	515	2013-09-04 13:35:01 Wed	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4189	527	2013-07-22 22:48:56 Mon	1	http://10.60.149.179/	0	null	Paying2 正常
4187	527	2013-07-22 22:48:50 Mon	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4181	526	2013-07-22 22:48:27 Mon	1	http://10.60.149.179/	0	null	Paying2 正常
4179	526	2013-07-22 22:48:21 Mon	1	10.60.149.179/Cash/se	1	orderIdgrossstatusid:	Paying 正常
4173	525	2013-07-22 22:47:57 Mon	1	http://10.60.149.179/	0	null	Paying2 正常

图 10 软件行为监控

第二部分是平台用户行为监控可视化，这部分是对平台用户行为习惯监控数据的可视化，其子部分包含了多维度的用户网络行为信息，如用户上网时间段的分布、用户访问的网站类的成分等，通过多维度信息展现用户的行为习惯。部分界面如图 11 所示。

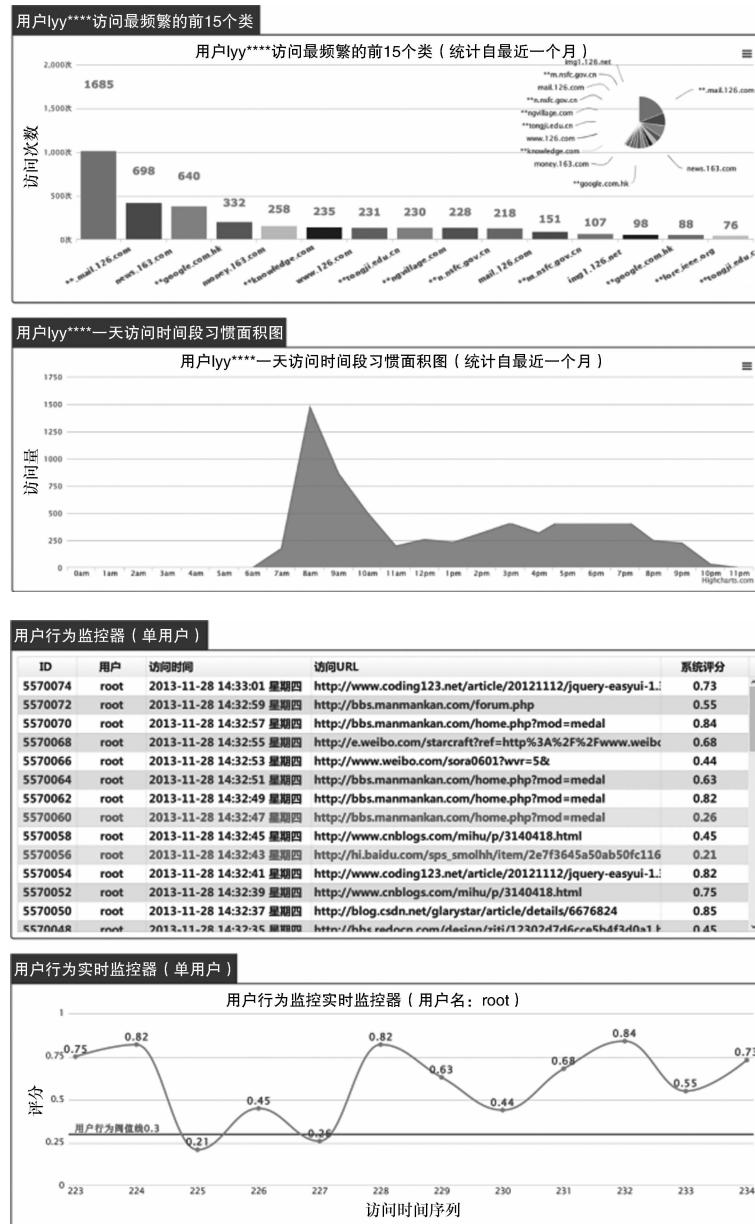


图 11 用户行为监控

第三部分为平台交易数据可视化，用于展示经过平台的交易数据，其数据是通过实时数据服务向受监控的外部电商平台获取的，包括全国交易量统计、实时交易量监控等信息，部分界面如图 12 所示。

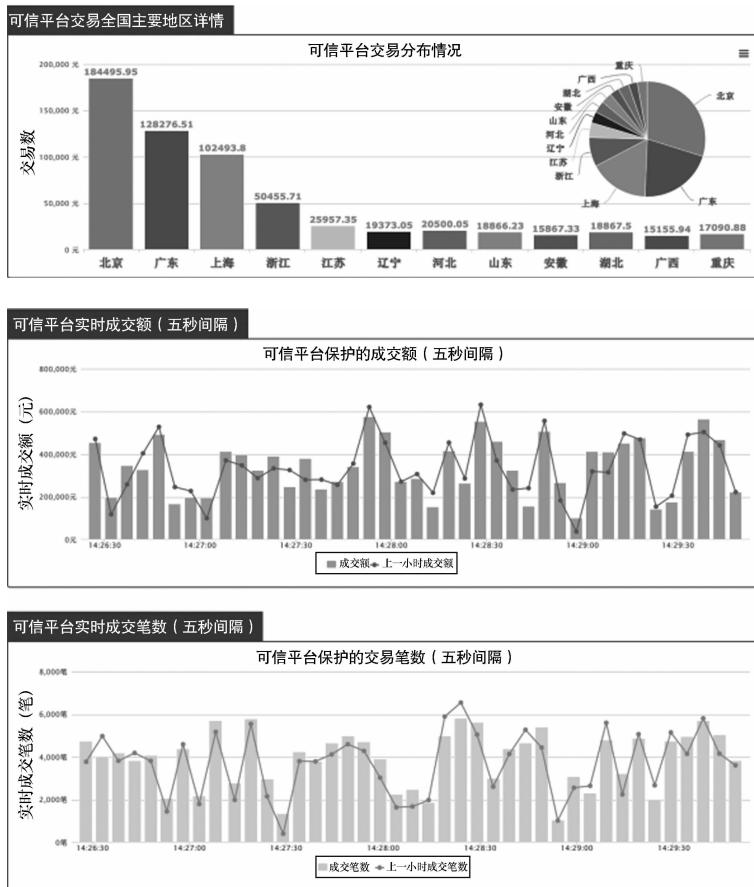


图 12 全国交易量监控

5.3 互联网医疗

在医疗领域，发展“互联网+”战略，需充分利用互联网、大数据等手段，构建医学影像、健康档案、检验报告、电子病历等医疗信息共享服务平台，积极利用移动互联网提供在线预约诊疗、候诊提醒、划价缴费、诊疗报告查询、药品配送等便捷服务，为医生及患者提供更优质贴心的服务，从而助力国家医改、促进医疗信息产业生态格局的形成。

前期工作基于 IPv6 下一代互联网技术，在 CNGI-CERNET2 主干网和校园网上完成医学科学教学和科研的信息服务平台的建设，利用 IPv6 及移动 IPv6 的优势为所有支持 IPv6 的高校用户、广大临床医护人员和医药界人士提供医学资讯信息、专业便捷的信息交流平台、丰富生动的医学科普知识，逐步使其建设成我国医学教育与科研信息资源的权威发布平台和资源共享平台。

下一代互联网医学教学科研平台应用示范系统组成功能如图 13 所示。即下一代互联

网医学教学科研平台应用示范系统由信息平台安全系统、教学课堂、医学科研协作环境、数字科普馆、基于 IPv6 传感器网络的医疗健康护理系统以及医学教学科研平台综合门户等 6 个功能子系统组成。

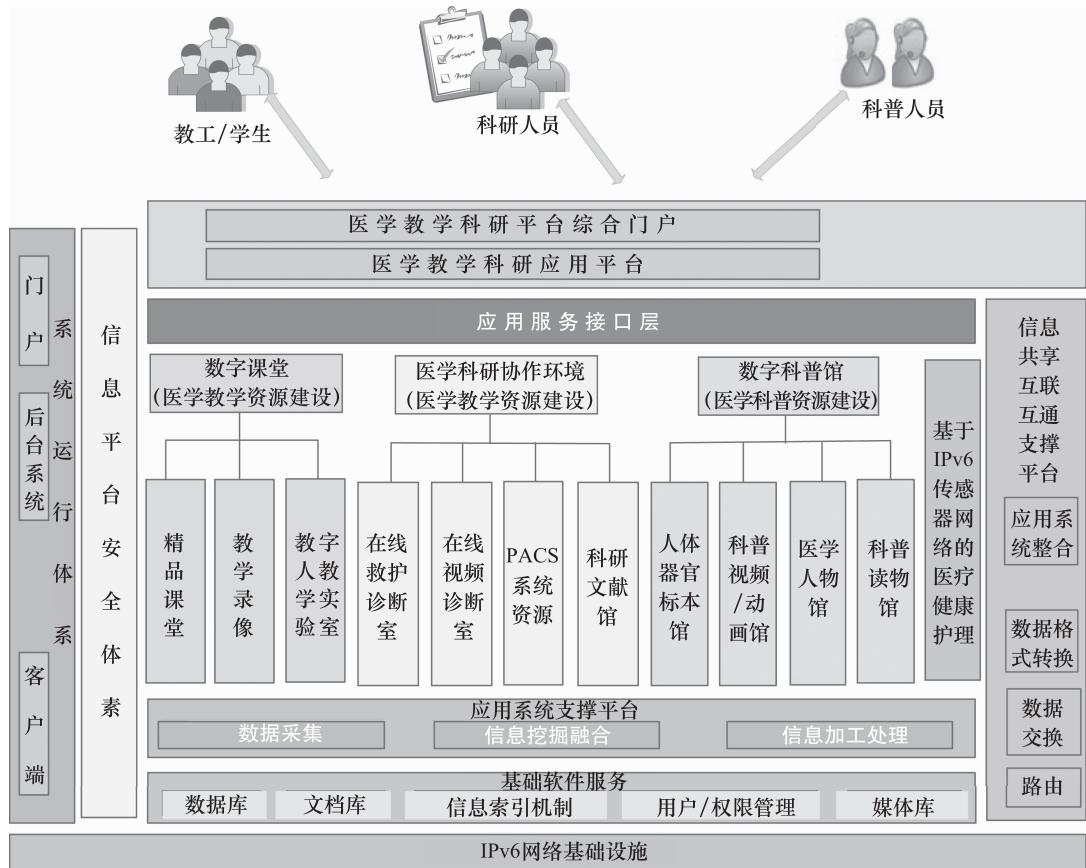


图 13 下一代互联网医学教学科研平台应用示范系统功能示意图

下一代互联网医学教学科研平台应用示范系统部署情况如图 14 所示。该系统构建在基于 IPv4/IPv6 双栈网络之上，覆盖上海、北京、武汉、杭州及汕头等 5 个城市。平台在同济大学建立了主节点，在北京大学人民医院、武汉大学医学部、浙江大学和华中科技大学建立了基于 IPv6/IPv4 双栈的分节点，在汕头大学组建了基于 IPv4 的分节点。在主节点组建了医学教学科研平台综合门户、医学科研精品课程，包含模拟人教学等视频录像、PACS 以及文献资料等资源库，在分节点组建了各节点所在单位各自的医学科研和科普资源库。系统配置了用于用户生命体征的实时采集、发送、统计及分析的基于 IPv6 无线传感器网络的医疗健康护理子系统。部分功能还包括医学教研论坛、真实患者 PACS 诊断学习案例建设、人体器官标本视频建设、脑电信号无线传感监测、在线视频诊断实时录像等。

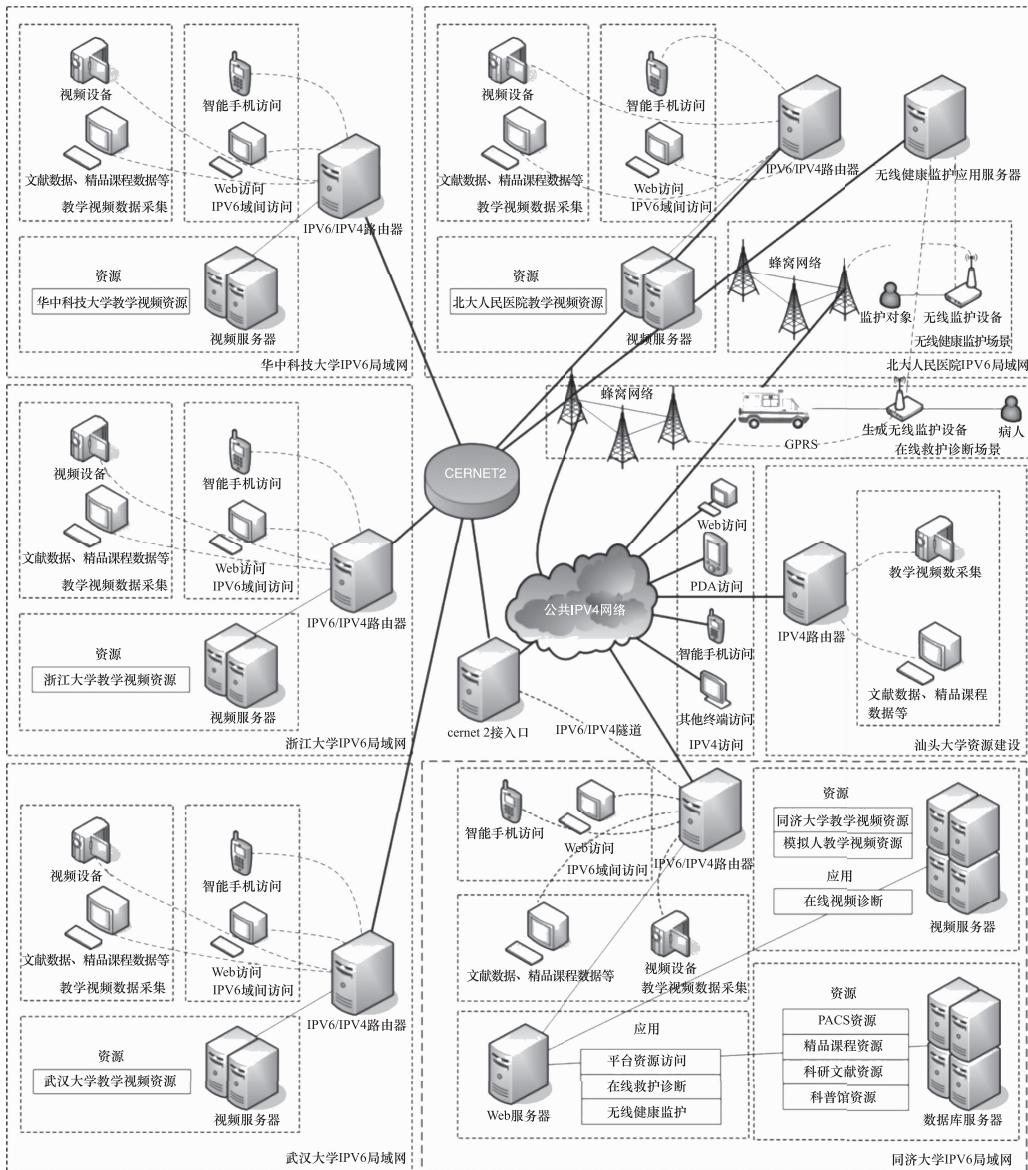


图 14 下一代互联网医学教学科研平台应用示范系统部署情况

6 结束语

在“互联网+”时代，数据已经渗透至各个行业，并且呈现出数量大、动态性、类型复杂等显著大数据特征，当前“互联网+”正全面应用到第三产业，形成诸如互联网金融、互联网交通、互联网医疗、互联网教育等新业态，同时不断向第一产业和第二产业渗透。本章分析和探讨国内外互联网行业发展现状，简要阐述大数据的研究现状与重

大意义，探讨大型数据资源服务平台架构，指出“互联网+”的发展思路与对策，并介绍了在智能交通和可信网络金融交易两大领域所开展的研究与应用工作。

参考文献

- [1] DC Digital Universe. Extracting Value from Chaos[R]. EMC, 2011-06.
- [2] Mary Meeker. Code Conference[J/OL]. 2015-05. <http://www.kpcb.com/internet-trends>.
- [3] Marco Annunziata, Peter Evans. The Industrial Internet: Pushing the Boundaries of Men and Machines, Available[J/OL]. http://www.ge.com/docs/chapters/Industrial_Internet.pdf.
- [4] 曹淑敏. 工业互联网和智能制造中、美、德比较及中国优势[R]. 中国信息经济年会暨中国信息化百人会 2015 年会, 2015-02-07.
- [5] Nature. Big Data[EB/OL]. 2013-03-20. <http://www.nature.com/news/specials/bigdata/index.html>.
- [6] Tony Hey. The Fourth Paradigm: Data-Intensive Scientific Discovery[R]. Microsoft Research, 2009-10-16.
- [7] Science. Special Online Collection: Dealing with Data [EB/OL]. 2013-03-20. <http://www.sciencemag.org/site/special/data/>.
- [8] Big Data. ERCIM News[N]. 2012(89).
- [9] World Economic Forum. Big Data, Big Impact: New Possibilities for International Development[J/OL]. 2013-03-20. http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBig_Impact_Briefing_2012.pdf.
- [10] Big Data across the Federal Government [EB/OL]. 2013-03-20. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_1.pdf.
- [11] 李国杰, 程学旗. 大数据研究:未来科技及经济社会发展的重大战略领域[J]. 中国科学院院刊, 2012, 27(6):647-657.
- [12] A Global Pulse White Paper, Big Data for Development: Challenges & Opportunities[R]. Global Pulse, 2012-05-29.
- [13] 赛迪网. 大数据产业生态战略研究 2012[EB/OL]. 2013-03-20. <http://www.ccidconsulting.com/portal/bps/webinfo/2012/06/1338946052018270.htm>.
- [14] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large[R].
- [15] Clusters. Sixth Symposium on Operating System Design and Implementation[R]. San Francisco, 2004-09.
- [16] Alexander Hall, Olaf Bachmann, Robert Busow, et al. Processing a Trillion Cells per Mouse Click[J]. Proceedings of the VLDB Endowment, 2012, 5(11):1436-1446.
- [17] Joel Wolf, Andrey Balmin, Deepak Rajan. On the optimization of schedules for MapReduce workloads in the presence of shared scans[J]. The VLDB Journal, 2012, 21:589-609.
- [18] Jingren, Nicolas Bruno, Ming-Chuan Wu. SCOPE: parallel databases meet MapReduce[J]. The VLDB Journal, 2012, 21:611-636.
- [19] Yingyi Bu, Bill Howe, Magdalena Balazinska. The HaLoop approach to large-scale iterative data analysis [J]. The VLDB Journal, 2012, 21:169-190.
- [20] Autoid Labs homepage[J/OL]. <http://www.autoidlabs.org/>.
- [21] International Telecommunication Union, Internet Reports 2005: The Internet of things[R]. Geneva: ITU, 2005.
- [22] 孙其博, 刘杰, 黎彝等. 物联网:概念、架构与关键技术研究综述[J]. 北京邮电大学学报, 2010, 33

- (3):1-9.
- [23] 张亚勤. 与云共舞——微软云计算的新进展[R]. 中国计算机用户, 2009, 12-13.
 - [24] 成都物联网产业发展联盟. 物联网云计算信息动态 2012[R].
 - [25] 赛迪顾问. 中国政府云计算应用战略研究 2012[R]. 2012-05-18.
 - [26] 百度开发者中心. 百度开发者中心[J/OL]. <http://developer.baidu.com/service>.
 - [27] 腾讯开放平台. 腾讯开放平台[J/OL]. <http://wiki.open.qq.com/wiki/%E4%BA%91%E6%9C%8D%E5%8A%A1>.
 - [28] 阿里云. 阿里云[J/OL]. <http://www.aliyun.com/>.
 - [29] 万网. 万网[J/OL]. <http://open.www.net.cn/>.
 - [30] Jane W. S. Liu. Real-time Systems, First Edition [M]. Pearson Education North Asia Limited and Higher Education Press, 2002.
 - [31] 陈艳. 并发实时系统的模型及其形式化[D]. 广西师范大学硕士学位论文, 2008.
 - [32] Adam T L, Chandy K M, Dickson J. A comparison of list scheduling for parallel processing systems [J]. Communications of the ACM, 1974, 17:685.
 - [33] Hwang J J, Chow Y C, F D, et al. Scheduling precedence graphs in systems with inter-processor times [J]. SIAM Journal on Computing, 1989, 18(2):244.
 - [34] SihG C, Lee EA. A compile-time scheduling heuristic for interconnection constrained heterogeneous processor architectures[J]. IEEE Transactions on Parallel and Distributed Systems, 1993, 4(2):75.
 - [35] Ahmad I, Kwok Y K. On exploiting task duplication in parallel programs scheduling [J]. IEEE Transactions on Parallel and Distributed Systems, 1998, 9(9):872.
 - [36] Kwok Y K, Ahmad I. Static scheduling algorithms for allocating directed task graphs to multi processors [J]. ACM Computing Surveys, 1989, 31(4):406.
 - [37] Sarkar V. Partitioning and scheduling parallel programs for multiprocessors[M]. Cambridge:MIT Press, 1989.
 - [38] Kwok Y K, Ahmad I. Dynamic critical path scheduling: an effective technique for allocating task graphs onto multiprocessors[J]. IEEE Transactions on Parallel and Distributed Systems, 1996, 7(5):506.
 - [39] 郝东, 蒋昌俊, 林琳. 基于 Petri 网与 GA 算法的 FMS 调度优化[J]. 计算机学报, 2005, 28(2): 201-208.
 - [40] 杜晓丽, 蒋昌俊, 徐国荣, 等. 一种基于模糊聚类的网格 DAG 任务图调度算法[J]. 软件学报, 2006, 17(11):2277-2288.
 - [41] 杜晓丽, 王俊丽, 蒋昌俊. 异构环境下基于松弛标记法的任务调度[J]. 自动化学报, 2007, 33(6): 615-621.
 - [42] I Stoica, R Morris, D Karger, et al. Chord: A scalable peer-to-peer lookup service for Internet applications [R]. In Proc. of ACM SIGCOMM'01, SanDiego, CA, 2001-08.
 - [43] I Foster, CKesselman, Write, Jin H, et al. The Grid 2: Blueprint for a New Computing Infrastructure [M]. San Francisco:Morgan Kaufmann Publishers, 2004.
 - [44] R Berbner, M Spahn, N Repp, et al. WSQoS—A QoS Architecture for Web Service Workflows[R]. In Proc. of the 5th international conference on Service-Oriented Computing, 2007-09.
 - [45] A Weiss. Computing in the clouds[M]. Source Net Worker Publisher, ACM, New York, NY, USA, 2008.
 - [46] JiangCJ, SunHC, DingZJ, et al. An Indexing Network: Model and Applications[J]. IEEE Transactions on Systems, Man and Cybernetics: Systems, 2014, 44(12):1633-1648.
 - [47] 中国互联网络信息中心. 第 35 次中国互联网络发展状况统计报告[R]. 2015-02-03. http://www.cnnic.net.cn/hlwfzyj/hlwzbg/hlwtjbg/201502/t20150203_51634.htm.

- [48] 蒋昌俊. 大数据的勘探与分析的若干思考[R]. 国家自然科学基金委双清论坛报告, 2013-17.
- [49] 蒋昌俊. 互联网非合作环境下大数据的探析问题[R]. 中国科学院科学与技术前沿论坛报告, 2013.
- [50] Jiang C J, Ding Z J, Wang P W. An Indexing Network Model for information services and its applications [J]. In: Proceedings of the 6th IEEE International Conference on Service Oriented Computing and Applications, 2013, 290-297.
- [51] Deng X D, Jiang M, Sun H C, et al. A novel information search and recommendation services platform based on an Indexing Network[J]. In: Proceedings of the 6th IEEE International Conference on Service Oriented Computing and Applications, 2013, 194-197.
- [52] Du H, Wang J, Liu Y N. A time sequence protocol to achieve the effect of fair exchange without trusted third party[J]. Chin. Sci. Bull., 2014, 59:699-702.
- [53] Yu W Y, Yan C G, Ding Z J, et al. Modeling and monitoring of online shopping business processes based on system behavior patterns[J]. Journal of Computer Information Systems, 2013, 9: 1-8.
- [54] Jiang C J, Ding Z J, Wang J L, et al. Big data resource service platform for the internet financialindustry [J]. Chin. Sci. Bull2014, 59(35):5051-5058.

作者简介

蒋昌俊 博士, CCF 理事兼 Petri 网专委会主任, 同济大学教授。主要研究方向为并发理论、服务计算、网络信息服务等。

E-mail: cjiang@tongji.edu.cn。



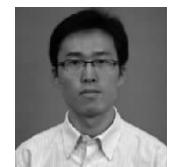
王俊丽 博士, 同济大学副研究员。主要研究方向为服务计算等。

E-mail: junliwang@tongji.edu.cn。



王鹏伟 博士, 东华大学讲师, 主要研究方向为服务计算、网络信息服务等。

E-mail: pwei.wang@gmail.com。



丁志军 博士, CCF 高级会员兼 Petri 网专委会秘书长, 同济大学教授。主要研究方向为服务计算、Petri 网理论及其应用等。

E-mail: dingzj@tongji.edu.cn。



基础计算系统可靠性研究进展

容错计算专委会

摘要

随着信息技术的发展，基础计算系统逐渐从传统的分布式计算系统和并行计算系统发展为网格计算系统，并最终衍生出云计算系统。各行各业对其依赖程度也不断加深，特别是在金融、电信、交通等一些关系国家安全与国计民生的关键领域，基础计算系统更是发挥着举足轻重的作用。开展面向基础计算系统的可靠性研究，有助于提高基础计算系统的可靠性，对于保障国民经济平稳运行具有重要意义。鉴于现代基础计算系统中云计算平台和互联网络的核心地位，本报告将综述国内外云计算平台和互联网络可靠性研究现状和进展，重点分析云计算平台中的多故障模式可靠性评估技术、故障注入技术、软件可靠性技术和硬件组件维护策略，以及互联网络中的网络可靠性评估技术和可靠性参数优化技术。

关键词：基础计算系统，可靠性，云计算，网络可靠性，软件可靠性，维修策略，故障注入

Abstract

With the development of information technology, traditional computing infrastructures, as distributed computing systems and parallel computing systems, have evolved into grid computing systems, and finally cloud computing systems. Modern computing infrastructures are very important in our society, and they are playing a pivotal role in those critical industries such as finance, telecommunication, transportation. Thus, the research on reliability of computing infrastructures is important because it can help to improve infrastructures' reliability and make them work better. Cloud computing systems and interconnect network are the base of modern computing infrastructures. This report provides an overview of state of the art, recent and expected progress, and open questions in reliability research of cloud computing systems and interconnect network. It covers reliability evaluation techniques for multiple failure modes in cloud computing systems, fault injection technology, software reliability and hardware maintenance strategies, as well as reliability assessment of interconnect network and reliability parameters optimization techniques.

Keywords: Computing infrastructure, reliability, cloud computing, network reliability, software reliability, maintenance policy, fault injection

1 引言

随着信息技术的发展，基础计算系统逐渐从传统的分布式计算系统（distributed computing system）和并行计算系统（parallel computing system）发展为网格计算系统（grid computing system），并最终衍生出云计算系统（cloud computing system）。各行各业

对其依赖程度也不断加深，特别是在金融、电信、交通等一些关系国家安全与国计民生的关键领域，基础计算系统更是发挥着举足轻重的作用。

应用于金融业、电信业、能源、交通、航空等关系到国家经济安全和社会安全的关键领域的基础计算系统不仅要具有极强的信息处理能力，还要具有较高的可靠性，能够提供高速、稳定、持续的信息处理服务。在上述关键行业应用中，基础计算系统发生延误和失效可能会造成难以估量的损失。表1列出了2008年以来全球范围内部分有重要影响的基础计算系统失效事件及失效原因，这些事件大多造成了难以估量的经济效益损失和社会效益损失。据Qualix Group权威数据统计，运输业发生1分钟的系统失效事件平均会损失15美元，银行业发生1分钟的系统失效事件平均会损失27万美元，通信业发生1分钟的系统失效事件平均会损失35万美元，制造业发生1分钟的系统失效事件平均会损失42万美元，而证券业发生1分钟的系统失效事件平均会损失高达45万美元。因此，开展面向基础计算系统的可靠性研究，有助于提高基础计算系统的可靠性，对于保障国民经济平稳运行具有重要意义。

表1 2008年以来全球范围内部分有重要影响的基础计算系统失效事件及失效原因

公司	时间	失效事件	失效原因
伦敦证交所	2008.9	被迫暂停交易7小时造成重大经济损失	交易量激增导致交易系统拥塞，进而大面积崩溃
Gmail	2009.2	全球性故障	系统更新导致资源过载，引发全球性的断线
微软	2009.3	云计算平台 Azure 宕机	中心处理和存储设备故障
Salesforce	2010.1	Salesforce 宕机超过1小时	数据中心的“系统性错误”
Terremark	2010.3	宕机7小时，导致迈阿密数据中心 vCloud Express 服务中断	Terremark 失去连接
微软	2010.9	连续出现至少3次 BPOS 托管服务中断	缺省配置错误
上海证交所	2010.1	交易业务主机宕机，导致 ETF 业务被迫暂停3小时	业务向备用主机迁移过程中订单处理能力大幅下降
澳洲国家银行	2010.11	数百万储户3天无法提款，甚至影响了澳洲汇丰和澳洲花旗的结算	档案受损，大量客户交易数据被错误清除
亚马逊	2011.4	云数据中心服务器大面积宕机	迁移策略的设计缺陷
东京证交所	2012.2	导致241种股票现货交易被迫暂停3小时	信息传输系统发生故障
东京证交所	2012.8	所有金融衍生商品交易被迫暂停1.5小时	交易主机故障

鉴于现代基础计算系统中云计算平台和互联网络的核心地位，本报告将综述国际上云计算平台和互联网络可靠性研究现状和进展，重点分析云计算平台中的多故障模式可靠性评估技术、故障注入技术、软件可靠性技术、系统故障特征和硬件组件维护策略研究，以及互联网络中的网络可靠性评估技术和可靠性参数优化技术研究。

以云计算为代表的现代基础计算系统采用了新的计算理念和新的资源交付方式。在计算理念上，将计算通过Internet交给计算平台来处理；在资源交付上，将IT资源、系

统资源和应用等整合为服务提供给用户^[1]。

典型的现代基础计算系统架构如图 1 所示。下面分别对该架构图中的云计算平台、异构互连网络和多样化终端三个层次做进一步介绍^[2]。



图 1 典型的现代基础计算系统架构

(1) 云计算平台

云计算平台有两个方面的用途。一是提供大规模的计算资源，为大数据的存储和计算提供基础设施，这项功能和分布式系统类似，即利用多台机器完成一台普通机器无法完成的任务。二是为大量用户提供瘦客户端的服务，即将计算和存储的部分移到远端而非客户端，这就要求云平台能为每个用户提供私人空间，同时不能造成资源浪费，在这一功能的实现中，虚拟机发挥着重要的作用。

典型的云计算平台如图 2 所示，由管理服务器集群、数据库集群、主机集群和辅助存储四部分构成。管理服务器集群是云平台的管理模块，管理并协调整个云平台的正常工作。数据库集群主要用来协助管理云环境的资源。主机集群是云环境中对外提供服务的最重要的模块，由 Hypervisor 集群实现。辅助存储一般与管理模块和每个主机都相连，连接方式包括 NFS 和 iSCSI。

目前，有很多开源的云计算平台搭建软件，如 CloudStack、Eucalyptus 和 OpenStack，它们都有一些共同的特点，如不依赖于某个特定的 Hypervisor、有虚拟网络的支持等，但相比之下，CloudStack 有更多更高级的特性，比如支持快照，支持虚拟机在主机之间的动态迁移等。

(2) 异构互连网络

早期的基础计算系统互连网络不仅速率慢而且技术体系非常复杂，设备“七国八制”，物理层和链路层曾经出现的技术就有串口通信、ATM、以太网、xDSL 等，以太网技术开始的定位是局域网技术的一种，广域网则普遍使用的是 E1\SDH\ATM 等技术，也有不少使用的是远程拨号技术。

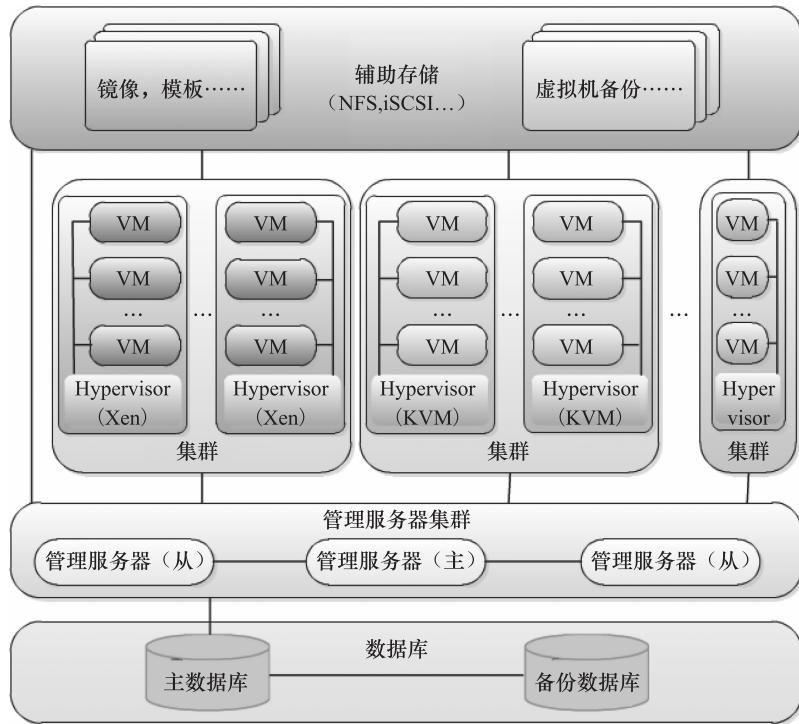


图 2 一个典型的云计算平台

经过二十多年的发展，以太网技术从局域网技术发展为城域网技术，并逐步走向广域网技术，成为唯一一种实际组建整个有线网络的技术。有线互联网以其特有的方式完成了技术的统一，在这期间，TCP/IP 协议族也从 ISO 协议族、Banyan 协议族、AppleTalk 协议族、Novell 协议族、IBM 协议族等众多的协议族中脱颖而出，最终成为互联网的标准协议，有线网络技术实现了一体化。

近年来，无线网络技术的发展则更是为 Internet 的进一步扩展提供了有力的支持。同时无线接入技术也获得了长足的进步，以 IEEE802.11 (WLAN) 系列、GSM、GPRS、802.15 (Bluetooth) 系列、802.14 (Zigbee) 系列、802.16 系列 (WiMAX)、3G/4G 等为代表的无线标准和相应接入技术相继出现并发展，为扩大网络覆盖区域、降低网络成本和简化网络部署提供了有效的解决方法。基于 IEEE802.11b/g 的 WLAN 和正在研究发展之中的无线 mesh 网络 (wireless mesh network, WMN) 是有可能推广为实际应用的下一代无线接入网及无线 - 有线结合网络的接入技术趋势。

(3) 多样化终端

现代基础计算系统对终端的基本要求就是有稳定的网络连接和基本的计算能力，必须适应不同用户群和云计算发展的要求，以下按不同客户群对相应的云终端的发展趋势进行分析^[3]。

1) 企业用户。适应绿色 IT，节能降耗，降低成本，满足差旅对便携性的要求。网络终端可以适应大部分使用常规应用的企业，但为了适应企业员工差旅的要求，仍需向提

高便携性方向发展，解决使用方面的局限性问题。

2) 个人用户。上网本、UMPC、MID 等目前存在续航能力差、便携性不足的问题，因此未来应该朝着便携、方便、易用的 UI 人机界面方向发展。其中便携性涵盖了节电、优化电源管理、快速启动、支持 3G 网络、小尺寸、免升级、免维护、高度的安全可靠性。

智能手机存在屏幕小、输入不便的问题，需要向稍大屏幕、外接屏幕、创新输入模式的方向发展。便携的折叠式屏幕、眼镜式屏幕等新型产品陆续面世，通过手势、视觉输入的创意也不断涌现，从而解决了这两个问题，为智能手机在云技术领域的应用扫清了障碍。

考虑到个人用户的应用场景，这种设备还应该满足用户在音乐、视频等方面的需求。终端设备自带屏幕只需支持半高清视频，但外接到大屏幕显示器和电视时，应该能够支持全高清视频的播放。当外接到大屏幕显示器时，用户应该可以通过外接输入设备（如无线键鼠）使用云计算应用软件和其他终端自带的功能（如视频、声音播放等）。

3) 家庭客户。离子平台作为小尺寸的固定终端，应具备一定的处理能力，能播放高清电影，具备 3D 处理芯片和宽带上网功能。通过这样的终端，用户可以看电视、看网络电影、玩游戏，也可以使用云平台上的软件进行移动办公。离子平台目前仍存在对操作系统的维护要求和安全风险，考虑到家庭客户的应用场景，需要固化操作系统并家电化。

2 国际研究现状

国际上针对以云计算为代表的现代基础计算系统的可靠性研究已经获得了丰富的研究成果，本节从多故障模式的云计算平台可靠性评估、基于故障注入的云计算平台可靠性评测、云计算平台的开源软件可靠性评估、系统部件维护策略和互联网络可靠性评估五个方面进行如下综述。

2.1 考虑多故障模式的云计算平台可靠性评估国际研究现状

云计算平台作为大规模基础设施，与传统服务器系统相比，其可靠性行为呈现出各种故障模式，主要包括以下三种。一是动态故障：随着计算部件集成化和智能化的加强，尤其是各种冗余机制和容错机制的应用，使得系统呈现出多种动态可靠性行为，包括功能依赖行为（Functional Dependency, FDEP）、冷备份机制（Cold Spares, CSP）、优先级失效（Priority AND, PAND）、顺序相关失效（Sequence Enforcing, SEQ）等动态可靠性行为。二是共模故障：典型的共模故障大致包括三种类型，①不可预测的自然灾害，如地震、台风、水灾、雷电、火灾等；②基础设施的损坏，如建筑物倒塌、电源中断等；③操作失误，如误操作、人为蓄意破坏等。共模故障的存在增大了系统中各部件的联合失效概率，使得系统中冗余容错机制的有效性降低，造成系统局部或整体失效。三是传

播故障：由于系统内某个部件的失效而导致系统其他部件同时失效的可靠性行为称为传播故障，如当路由器或交换机发生故障时，其所管理的整个子网的网络通信就会失效。

国际上在多故障模式计算系统可靠性评估方面的研究较为集中，除了部分欧洲学者，大量研究工作集中于美国 Duke University 的 Kishor S. Trivedi 教授领导的研究团队、美国 University of Virginia 的 Joanne Bechta Dugan 教授领导的研究团队和美国 University of Massachusetts Dartmouth 的 Liudong Xing 教授领导的研究团队。

下面针对国外在多故障模式计算系统可靠性评估方面所取得的研究成果，从动态故障系统可靠性评估、共模故障系统可靠性评估和传播故障系统可靠性评估三个方面进行如下综述。

(1) 动态故障系统可靠性评估

基于静态故障树的传统可靠性评估方法专注于静态逻辑或静态故障机理，不能有效处理动态故障系统中包含的时序相关失效、备份冗余机制、功能依赖机制等动态可靠性行为。另一方面，基于 CTMC 的状态空间方法只能处理指数分布失效，同时存在严重的状态爆炸问题，即使对于一个简单的系统，CTMC 模型的建立和求解也是非常繁琐的，有时甚至由于运算量巨大而无法求解^[4-5]。

自 1992 年开始，Dugan 等研究人员通过引入表征动态故障特性的动态门，建立了动态故障树模型，并给出了处理动态故障树的三种模块化方法，避免了将整个动态故障树转化成 CTMC 模型而导致的状态爆炸问题。在后续的研究中，Dugan 领导的研究小组开发了 DIFTree 和 Galileo 两款动态故障树分析软件，大大强化了动态故障树方法在工业界的应用^[6-7]。研究表明，动态故障树法综合了组合分析和状态空间分析两者的优点，是解决动态故障系统可靠性评估问题的有效途径。在 Dugan 研究的基础上，Huang 和 Chang 提出了细粒度的模块化方法，使得模块化操作能够深入到动态门子树中，进一步提升了动态故障树分析方法的效率^[8-9]。

(2) 共模故障系统可靠性评估

共模故障会明显增加系统联合失效概率，并对系统最终可靠度具有重要影响，因此得到了广泛和深入的研究。在静态故障和共模故障共存情况下，存在两类分析方法——显式方法和隐式方法。在显式方法中，系统故障树模型包含刻画共模故障的基本事件^[10]；而隐式方法则采用两阶段处理策略，先构造不含共模故障的系统模型和可靠性公式，然后采用后处理的方法在计算最终可靠度时考虑共模故障^[11]。当进一步考虑动态故障和共模故障共存时，系统可靠性评估更加复杂。Xing 领导的研究团队通过引入新的动态门来刻画共模故障，并基于动态故障树进行系统可靠性建模，然后采用分解和聚合（Decomposition and Aggregation, EDA）方法进行可靠性评估^[12-13]。

(3) 传播故障系统可靠性评估

由于系统内某个部件的失效而导致系统其他部件同时失效的可靠性行为称为传播故障。传统的不完全覆盖研究是传播故障研究的一个重要组成部分。在分布式计算系统中，由于病毒和恶意攻击的广泛存在使得传播故障研究显得更为重要。

早期的传播故障系统可靠性评估方法主要包括状态空间分析方法和组合分析方法。

由于分析能力的局限性，这些方法只能应用于较小规模的传播故障系统。为此，研究人员根据可靠性理论的新进展不断提出新的分析方法，其中比较典型的有 Levitin 等人提出的基于发生函数的分析方法^[14-15]。该分析方法首先基于 U 函数计算各个系统部件的概率质量函数，然后利用组合操作构造整个系统的 U 函数描述，最后根据系统 U 函数计算获得系统可靠性指标。发生函数方法比较适合具有规则结构的传播故障系统。

近年来，研究人员针对具有一般结构的传播故障系统可靠性展开广泛研究。Amari 和 Xing 等提出了 SEA 分析方法（Simple and Efficient Algorithm, SEA）^[16-17]，该方法利用不完全覆盖故障的全局传播性质，通过全概率公式把原问题分解为不完全覆盖故障发生和不发生这两个互斥子问题，然后进行子问题求解和结果综合。在此基础上，Chang 和 Myers 等人把决策图分析方法引入子问题求解过程，把子问题对应的故障树转化为等价 BDD 并进行快速分析，从而提升了 SEA 方法的分析能力^[18-19]。针对二值状态变量数量和二值决策图尺度爆炸问题，Xing 领导的研究团队进一步根据不完全覆盖传播故障和局部故障之间的互斥性，采用三值决策图代替二值决策图进行分析，并根据单点失效特性设计了相应的三值决策图操作方法，实验数据表明三值决策图方法能够明显提升可分析传播故障系统的规模^[20-21]。

2.2 基于故障注入的云计算平台可靠性评测国际研究现状

故障注入技术首次在国际上提出是在 20 世纪 70 年代，开始仅应用于工业界对容错系统的设计和验证。到 80 年代中期，作为系统中容错机制的实验评价方法，故障注入技术开始被科研部门和高校普遍采用。随着研究的深入及技术的成熟，进入 90 年代后，该技术引起越来越多的研究人员和设计者的重视，其应用也随之广泛起来。

国外对故障注入的研究起步较早，对故障注入理论的研究已相对深入。如美国的 California、CMU、Duke、IBM、Illinois、Michigan、NASA、RS、Tandem、Texas A&M、Virginia 以及法国的 LAAS-CNRS 等大学和研究部门，他们的故障注入研究工作都处于国际领先地位，他们成功研发的许多功能强大的故障注入工具已投入使用，并取得了良好的效果。此外，瑞典的 Chalmers 大学、奥地利的 Vienna 技术大学、德国的 Dortmund、Erlangen-Nürnberg、Karlsruhe 等大学有许多科研人员从事故障注入研究方面的工作，而英国、日本、意大利、澳大利亚等国的一些高校和研究部门也在从事故障注入研究。针对传统单机服务器的故障注入工具已经有了很长时间的发展，并且有成熟的工具出现。例如 FIAT 工具，其由美国 CMU 大学开发，通过修改进程的内存镜像向目标机注入内存故障、通信故障等。还有 FERRARI 工具，其由美国德克萨斯大学奥斯汀分校开发，通过 Linux 系统自带的 PTrace 调试接口，利用软件陷阱触发故障注入，可以向系统的 CPU、内存、总线等部件注入故障。

为了满足对虚拟化云计算平台可靠性的评测要求，Michael Le 等对在虚拟机中进行故障注入的困难和相应方法进行了综合概括的论述^[22]。文中提到研究虚拟机故障注入的重要性：首先用虚拟化故障注入的方法可以促进对非虚拟化系统容错能力的研究，其次用

故障注入的方法可以评测虚拟化系统的可靠性。它以 Xen 作为实例，通过 VMM 和 VM 向半虚拟化和全虚拟化的机器注入大量的代码故障、内存故障和寄存器故障等。文中提出了通过在 VMM 中修改页表实现内存故障注入的方法，也提到了如何在 VMM 中修改多个 VM 的资源引用列表评测 VM 独立性，还实现了通过与其他虚拟机依靠共享内存来记录故障注入结果和日志的方法。Sankalp Agarwal 以 Xen 框架为例，利用检查点回卷，提供了一套用户态的检查点库，包括基本的检查点回卷算法（如同步、异步算法等），通过此库能够使多个 VM 互相之间就检查点信息进行交换。

2.3 云计算平台的开源软件可靠性评估研究国际进展

可靠性是衡量所有软件系统最重要的特征之一。不可靠的软件会让用户付出更多的时间和金钱，也会使开发人员名誉扫地。IEEE 把软件可靠性定义为在规定条件下和规定时间内软件不发生失效的概率。该概率是软件输入和系统输出的函数，也是软件中存在故障的函数，输入将确定是否会遇到所存在的故障。

云计算平台中广泛采用开源软件（OSS），也叫开放源代码软件（Open Source Software）。开源软件是一种源代码可以任意获取的计算机软件，这种软件的版权拥有者在软件协议的规定下保留一部分权利并允许使用者自由学习、修改、拷贝、散布（distribution），同时可以改善和提高开源软件的质量。目前著名的开源软件包括 Linux、Apache、Mozilla 和 MySQL 等。和传统闭源软件一样，软件可靠性是开源软件从发布、测试到运行的一项重要指标，这不但关系到软件产品质量，也牵涉到开源软件企业的行业声誉。

软件可靠性模型是软件可靠性研究中备受关注、成果最多、最活跃的一个领域。从 Hudson 的工作开始，到 1971 年 Jelinski-Moranda (J-M) 模型的发表，至今已公开发表了几百种模型。软件可靠性模型旨在根据软件失效数据，通过建模给出软件的可靠性估计值或预测值。它不仅是软件可靠性预计、分配、分析与评价的最强有力的工具，而且为改善软件质量提供了指南。

开源软件可靠性模型大致可分为三种。第一种模型是基于修改非齐次泊松分布（NHPP）的可靠性模型，也就是在传统的闭源软件可靠性增长模型的基础上混合了开源软件的特点，并进行相应修改后得到的可靠性模型；第二种模型是基于随机微分方程的开源软件可靠性模型；第三种则是其他开源软件可靠性模型。

基于修改的 NHPP 开源软件可靠性模型主要是从经典的闭源软件 NHPP 的可靠性模型发展而来。成功的闭源软件 NHPP 的可靠性模型为其提供了可参考和借鉴的经验。例如 Goel 和 Okumoto^[23]提出的传统的 NHPP 可靠性模型，把软件故障检测看作服从非齐次泊松分布的随机过程。随后，Yamada 等人^[24]在研究了基于 NHPP 的可靠性增长模型后，对于平均值函数呈现的 S 形状，提出了基于修改的 NHPP 的 S 形状的可靠性增长模型。

虽然开源软件和闭源软件有很多不同之处，但是在故障检测上却存在诸多相似之处。可以把开源软件的故障检测看作服从非齐次泊松分布的随机过程，Li 等人^[26]在考虑了开

源软件的特点后提出，和闭源软件不同的是，开源软件从发布起就吸引了大量的使用者，其数量远远超过了闭源软件的测试者。但是随着时间的发展，使用者的数量达到最大值后开始下降，表现为驼峰（hump-shaped）模型。从使用 Apache 和 Gnome 数据集进行的开源软件可靠性测试可以知道，该模型准确覆盖了故障检测过程，用均值平方错误（mean squared error）标准进行检测，可靠性远远超过传统的 NHPP 模型和 S-shaped 模型。不过该模型并未考虑开源软件故障检测是一个无限的过程，随着时间的发展，可能检测出无穷个错误。基于此，又进一步提出了以下的基于随机微分方程的可靠性模型^[27]。但在实际中，开源软件的故障检测总数量受到版本发布时间和成本费用因素的综合作用，不可能将其视为无穷。所以，合理的测算开源软件的故障检测总数量对于软件可靠性的评估具有重要作用。

在参考了闭源软件的威尔布（Weibull）分布可靠性模型后，Zhou 和 Davis 提出了两个参数非线性威尔布分布回归模型^[25]。Tamura 和 Yamada 提出融合神经网络和对数泊松执行时间的软件可靠性模型，还提出了基于混沌理论和风险率变化的嵌入式开源软件可靠性模型^[28]。

2.4 系统部件维护策略国际研究进展

对于大型云计算平台而言，从部署完成之日起，系统的可靠性、可用性、安全性就完全依赖于合理的维护操作。在整个生命周期中，系统维护所占据的比例超过生产制造。对于提高系统可靠性而言，在维护操作上花费的人力和物力也必将会超过生产制造、系统构架、冗余配置等工作上的投入^[29]。

在过去的几十年中，学术界展开了大量以可靠性、经济性为目标的维护模型和维护策略研究，针对不同的系统失效模式和具体的维护约束条件，采用更换和预防性维护相结合的维护策略，以提高长期经济效益或系统可靠性为目标对维护策略进行优化。

按照系统模型类型将维护策略分为独立部件的维护策略和考虑部件相关性的维护策略。

（1）独立部件维护策略

在劣化系统（deteriorating system）中，随着工作时间的延长，系统性能逐渐下降，直到不能满足工作需要而失效。实际应用中面临的大部分系统均属于劣化系统。在目标部件年龄已知的情况下，合理的维护操作应该于劣化失效前对部件进行更换。

年龄更换策略中，若部件在 T 之前发生失效则对部件进行修复性更换，当部件达到年龄 T 仍然正常工作，则对部件进行预防性更换^[30]。指定的时间 T 称为计划内更换时间。该策略假设失效在发生后能够立刻得以检测，更换采用的是全新部件，并且立刻开始工作。年龄更换策略是一类广泛使用的维护策略，国内外很多学者对年龄更换策略展开了大量研究，主要分为一般性模型问题和面向不同的应用环境、系统结构和失效模式建立特定的维护模型并寻求最优更换年龄的问题。

年龄更换策略的设计目标针对单个部件，策略实施需要获得部件年龄（或者使用情

况) 的准确信息。针对此问题, 基于失效次数 N 的更换策略由于较易实施, 常作为年龄更换策略的替代策略, 并与不完美维护模型相结合, 对维护策略进行评价。基于失效次数 N 的更换策略在前 $N-1$ 次失效中采取维修操作, 第 N 次失效采取更换。该策略适用于部件运行时间不易统计或运行中由于时间、费用问题不易进行部件更换的情况, 尤其是很多同类型部件构成的大型复杂设备。

一个新部件在 $t = 0$ 时开始工作, 部件失效能够立刻被检测到并更换为新部件, 同时, 在时间点 $kT(k = 1, 2, \dots)$ 上, 无论部件年龄是多少, 都采取更换操作, 该策略称为成批更换策略。对于成批更换策略而言, 更换操作在失效时以及周期时间点上进行, 此类策略的优势在于可操作性强, 无需保存过多失效时间相关的信息。但其缺陷也很明显, 对于经典的 BRP 策略, 一组工作部件将会分别在失效时或者在固定周期为 τ 的维护时间点上被更换, 在周期点上部件依然较新时, 更换操作会造成浪费。多数改进策略力求最大程度减少不必要的失效更换引起的费用。

对于单个部件, 在每次失效的时候进行最小维修使部件恢复工作, 同时进行周期的预防性更换, 该类策略称为周期更换策略。对于由大量元件组成的设备, 对单个元件的失效更换则相当于对整个设备的最小修。考虑多个部件组成的系统, 每个部件由大量元件组成, 部件失效发生时采取最小修, 在周期点上对所有部件进行预防性更换, 该策略为失效最小修的周期更换策略的多部件版本, 也可以视为分组更换策略的一种。

(2) 考虑部件相关性的维护策略

复杂系统大多由多个子系统构成, 考虑到子系统之间存在费用相关性、随机相关性和结构相关性, 面向此类系统的维护策略与单部件系统的维护策略有所不同, 一个子系统上优化的维护操作依赖于其他子系统的状态。例如, 若两个子系统在维护操作上存在费用相关性, 那么维护策略就需要统筹考虑, 同时维护两个子系统可能会比逐个维护费用更低, 一个子系统的失效可能带来维护另一个子系统的机会。

在很多实际系统中, 部件可以分为多个组, 每组由多个相似的部件组成。考虑到部件存在经济相关性, 在维护策略指定的时间点对一组部件进行更换的效果要好于分别对单个部件进行单独更换, 这一现象称为规模经济性 (economies of scale), 采用分组操作的维护策略在研究中也受到了广泛关注。研究中一类重要问题是如何对部件进行分组以便在发生失效的时候进行更换, 当部件组装和拆卸的费用不同时, 此类问题具有重要意义。另一类问题是通过在系统设计中加入冗余部分以降低费用, 还有类问题考虑为独立运行且具有相同失效分布的多个系统设计分组维护策略。按照问题的性质, 分组维护策略也主要分为三类。第一类称为基于年龄 T 的分组更换策略, 当系统到达年龄 T 的时候进行分组更换。第二类策略称为 m 失效分组更换策略, 当系统中的失效次数到达 m 后对系统进行更换。第三类策略为前两类的混合策略, 称为 (m, T) 分组更换策略, 当系统年龄到达 T 或者失效次数到达 m 两者之一先发生时即进行分组更换。

系统停机提供了很好的 CM 和 PM 机会, 基于系统停机机会的维护策略尤其适合于串联系统, 其中任意单个部件的失效都会引起系统宕机, 从而可以对其他部件进行预防性维护。文献 [31] 研究了多部件串联系统中的机会维护策略。在该类系统中, 当一个部

件进行 PM 的时候其他部件必须停止工作，从而提供了 PM 的机会。文中提出了基于动态规划的 PM 调度算法，优化目标为使短期累积的维护费用节约数最大。

2.5 互联网络可靠性评估国际研究进展

在基础计算系统中，网络变得越来越快捷、廉价、无处不在，随之而来的网络可靠性问题逐渐成为用户们关注的焦点，也给网络供应商、运营商带来严峻的挑战。网络的复杂性、动态性、多态性等特点使得传统成熟的可靠性评估方法，如可靠性框图（RBD）法、故障树分析（FTA）法等很难用于网络。近 10 年来，对网络可靠性的研究取得了很多成果，网络可靠性的内涵也由传统的基于网络拓扑结构的连通可靠性逐渐拓展到考虑网络流的容量可靠性，并向基于业务等考虑用户需求的性能可靠性延伸。网络可靠性评估主要分为 3 类^[32]：

1) 网络连通可靠性评估，指的是仅考虑网络拓扑结构，将“网络实现连通功能的概率”作为可靠性度量。

2) 网络容量可靠性评估，它在考虑网络是否连通的基础上，还考虑了网络中链路和节点的容量，将“存在满足一定流量需求的连通路径的概率”作为可靠性度量。

3) 网络性能可靠性评估，关注的是网络性能的动态变化对可靠性的影响，多以“某些性能参数不超过其规定阈值的概率”作为可靠性的度量。

近几年来还出现了以业务为中心的可靠性评估，综合考虑了网络的连通可靠性、容量可靠性和性能可靠性，将“网络对某业务的支持能力”作为业务可靠性的度量。

(1) 网络连通可靠性评估

1955 年，Lee 在“Analysis of Switching Networks”（交换网络分析）一文中，定义了以“能实现连通功能的概率”为度量的端可靠度，首次使用了以连通为规定功能的可靠性指标。连通可靠性也是最早提出的网络可靠性指标。研究将网络按照有无指定的源点分为有源网络与无源网络。有源网络可靠性指标分为 3 类：ST 可靠度（源点 s 与终点 t 保持连通的概率）；SK 可靠度（源点 s 与特定的端点集 K 保持连通的概率）；SAT 可靠度（源点 s 与网络中所有其他端点保持连通的概率）。无源网络可靠性指标也分为 3 类：两端可靠度（网络中两个端点间保持连通的概率）； k 端可靠度（网络中 k 个端点间保持连通的概率）；全端可靠度（网络中所有端点保持连通的概率）。

网络系统常用最小路集（或割集）描述。利用最小路集（或割集）计算网络可靠度时，首先要确定网络的最小路集（或割集），然后利用组合数学的容斥原理公式求网络的可靠度。一个路集对应网络的一个工作状态，一个割集对应网络的一个故障状态，因此该方法将网络可靠度表示为全部最小路集的并（或将网络故障度表示为全部最小割集的并），然后采用容斥原理去掉相容事件相交的部分，进而计算相应的可靠度。当最小路集（或割集）数为 m 时，其容斥原理公式包含 $2^m - 1$ 项。

采用容斥原理计算方法计算网络可靠度原理简单，易编程实现，在研究早期受到重视。然而当网络规模很大时将产生严重的组合爆炸问题，算法效率急剧下降。不交积和

算法的出现在一定程度上缓解了这一问题。不交积和法是运用不交积和定理计算网络可靠度的一种算法。将网络可靠度表示为全部最小路集的并，然后将并转化为不相交项的和，进而计算相应的可靠度。

(2) 网络容量可靠性评估

对于网络连通可靠性的评估是网络可靠性研究中最早开始进行的，因此研究成果很多。但在实际使用网络的过程中，会发现网络中不论是链路的容量还是节点的容量，都不是连通可靠性研究中所默认的“容量无限”。在网络连通的情况下，网络容量的限制依然会对“传输一定流量”的功能实现产生影响。因此，不仅关心网络中是否存在连通路径，还关心是否存在满足一定流量（物质、能量、信息）需求的连通路径。最早对这一问题进行研究的是美国普林斯顿大学的 Ford 教授等，他于 1956 年针对运输网、通信网、电网等一类容量有限的网络，基于图论提出了网络流模型，并于 1962 年首先给出了求解网络最大流的第一个算法——标号法，开拓了用数学网络理论来研究运输系统的思路，首次开始将链路容量与网络可靠性相结合。随后，出现了对路网能力可靠性的研究，即将路网能力可靠性定义为路段交通量不超过路段能力限制的概率。近年来，才开始对“定量信息通过网络的概率”（即流网络（Flow Network））的研究。首先对这一问题进行研究的是 K. K. Aggarwal 等，他们将一个系统是否故障定义为其能否成功地在源汇节点间传输要求的流量。为了简化问题，突出重点，他们做了如下假设：

- 1) 网络中节点无容量限制且完全可靠。
- 2) 网络中链路有容量限制，不能通过超出该容量信息流。
- 3) 链路只有故障？正常两种状态，故障时信息流无法通过。
- 4) 网络中链路故障的概率是统计独立的？

传统的网络流理论虽然考虑了网络容量的问题，但都是针对网络容量固定的网络。而现实世界中的网络系统受到多种不确定因素（如网络构件的降级运行、网络阻塞等）的影响，可能会导致网络拓扑结构、链路容量发生变化，从而表现出网络容量的随机性和多态性。所以，使用传统的网络流理论没有考虑到网络容量的这两个特性，用来解决随机环境下的网络实际问题已经不再合适。因此，近 10 年来，可靠性工作者致力于对流网络可靠性评估方法的进一步改善和对随机流网络的研究。

(3) 网络性能可靠性评估

虽然对于链路容量的可靠性评估方法已经考虑了网络流的负荷问题，但在实际的网络中，保证网络流量的路径并不像考虑链路容量的方法中所假设的那样自动依据拓扑结构生成（有路由算法参与其中），并且网络的拥塞、时延等故障也成为人们日渐关注的焦点。就 10 年来的研究成果来看，目前对于网络性能可靠性的评估还处于探索和尝试阶段，对于“网络性能可靠性”并没有如前两种可靠性那样明确的定义，而多以“某些性能参数（如吞吐量、丢包率、延迟时间等）不超过其规定阈值的概率”作为可靠性的度量。因此，网络性能可靠性研究需要解决的主要有以下 3 个难点：①如何针对网络的动态性、多态性等特点建立合理的可靠性分析数学模型？②基于网络性能的可靠性评估的指标是什么？③有哪些方法可以用于网络性能可靠性评估？围绕这 3 个问题，近 10 年来

的研究有以下进展。

1) 建立网络性能可靠性的数学模型。文献 [33] 提出了一个新的网络可靠性评估模型——流量路径模型。作者认为, 传统的概率图模型根本没有考虑网络的性能降级, 而概率容量模型不合理的假设 (如基于图论的最大流假设) 导致其只能通过网络拓扑结构来评估性能降级。作者提出的这一新模型并没有以拓扑信息为中心, 而是集中关注于流量路径的信息, 通过从流量路径到物理构件和容量的映射, 能简单明了地反映网络的性能降级情况, 从而能通过一种由上到下的方法评估通信网络的可靠性。

2) 网络性能可靠性的指标和解析算法。文献 [34] 提出了一个考虑网络及时可靠性评估的指标, 即以源信号在规定时间内到达终点的概率为该网络的及时可靠度。作者还提出了一种评估非循环传输网络 (Acyclic Transmission Network, ATN) 可靠性的新算法, 该算法基于使用状态枚举的扩展通用生成函数方法。同年, 文献 [35] 提出将路由缓冲区溢出的概率作为衡量网络拥塞的标准, 提出了一个基于状态空间的算法, 用于计算网络在 IP 层的性能可靠度。该算法假设网络中数据包的重选路由只由物理链路故障造成。

3) 网络性能可靠性的仿真方法。文献 [36] 提出了考虑网络拥塞的可靠性仿真方法。方法假设网络节点完全可靠, 链路可靠度为常数, 数据包到达时间服从泊松分布, 以成功数据包的数量与成功数据包及丢失数据包的数量之比作为网络的拥塞可靠度 (当存在数据包丢失时), 或者将成功数据包数量与仿真产生的所有数据包数量之比作为拥塞可靠度 (当没有数据包丢失时)。仿真采用事件驱动方法, 考虑 3 种事件: 数据包产生、数据包传输、数据包接收。先将时间清单初始化, 再按照所有事件发生的时间顺序将事件添加到事件清单上。该仿真考虑了 3 种路由选择算法: 更替路由法、泛洪法、静态最短路径法。

3 国内研究进展

国内关于以云计算为代表的现代基础计算系统的可靠性研究稍晚于国外, 但是跟踪和发展非常迅速。截至目前, 在基于 MDD 的全系统可靠性建模和分析、基于故障注入的计算平台可靠性评测、构件软件可靠性过程技术、基于 BDD 的互联网络可靠性分析、基于 BDD 的互联网络可靠性分析以及互联网络可靠性参数及优化等方面已经获得了重大研究进展。

3.1 基于 MDD 的全系统建模和分析国内研究进展

多值决策图 (Multiple-valued Decision Diagram, MDD) 是二值决策图 (Binary Decision Diagram, BDD) 的扩展形式, 相对于 BDD 具有更强的描述能力和分析能力。2009 年, Xing 和 Dai 首次把 MDD 引入系统可靠性评估领域, 提出了面向多状态系统可靠性的 MDD 分

析方法^[37]；在此基础上，2010 年，Amari 和 Xing 等把 MDD 分析方法扩展应用到多状态多性能级别计算系统可靠性评估中^[38]。

近年来，在国家自然基金的资助下，浙江师范大学可信计算团队在和国外主要研究团队交流合作研究的基础上，提出了面向承担阶段任务的基础计算系统可靠性的 MDD 分析方法，积累了丰富的研究经验，同时也获得不错的分析效果^[39-40]。在这些前期研究基础上，该研究团队针对更为复杂的动态故障、共模故障、传播故障共存的多故障模式云计算平台，提出了在 MDD 框架下多故障模式统一建模方法，以及基于单一系统级 MDD 模型的多故障模式统一分析方法，从而简化分析模型和分析方法的复杂性，扩大可分析系统的规模和复杂度、提升可靠性评估的准确性和效率^[41-42]。

图 3 直观说明了多故障模式云计算平台可靠性 MDD 分析方法研究中采用的研究路线。首先针对多故障模式云计算平台包含的各种故障模式建立相应的 MDD 建模方法，如动态门和相关重复事件的 MDD 建模、多个共模故障或多个传播故障的联合 MDD 建模、动态故障和传播故障的混合 MDD 建模；然后建立各种故障模式多值变量在形成正确排序时的约束条件，设计两阶段 MDD 模型生成方法和基于带备忘录 MDD 操作的快速 MDD 模型生成方法，用以处理共模故障/传播故障和依赖部件故障之间依赖关系，高效生成系统级 MDD 模型；进一步的，对于具有较大规模的系统实例，建立各种结构特征相关的启发式策略用以生成高性能的变量排序缓解 MDD 模型尺度爆炸问题，同时设计基于截断 MDD 的可靠度的上下限估计及其误差评价公式；最后基于典型多故障模式云计算平台实例的可靠性评估应用研究，比较和验证所提出的 MDD 分析方法的性能。

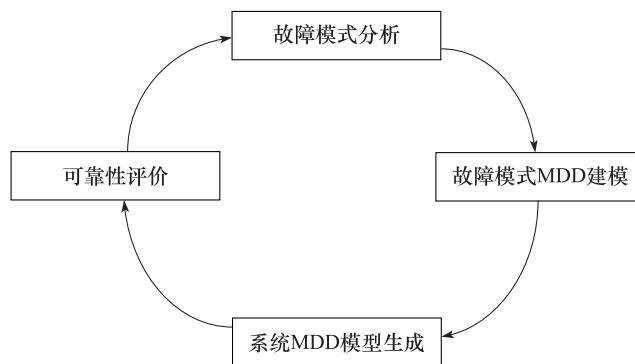


图 3 基于 MDD 的全系统建模和分析方法研究路线

针对多故障模式云计算平台可靠性评估这一重要问题，在 MDD 框架下的多故障模式的统一建模和分析方法，能够获得相对于 BDD 模型和状态空间模型更为简洁的决策图模型，能够缓解 SEA 方法和 EDA 方法的分析复杂性、实现分析自动化并提升分析效率。

3.2 基于故障注入的云计算平台可靠性评测国内研究进展

我国对故障注入的研究开始于 20 世纪 80 年代，首先是由哈尔滨工业大学杨孝宗教授领导的容错计算实验室（FTCL）开展，至今，此实验室在国内的故障注入研究方面仍

处于领先地位。在研究初期, FTCL 主要致力于硬件实现的故障注入方法的研究及芯片引脚级故障注入器的研制, 并于 1994 年开发完成容错性能评测设备 HFI-1 型故障注入器, 之后随着研究的深入不断对 HFI-1 进行改进, 形成了 HFI 系列故障注入器。此系列的故障注入器目前已应用于船舶总公司以及航天总公司下属部门研制的多种型号容错机中。与此同时, 该实验室不断扩展故障注入的研究领域, 在软件实现的故障注入方法、软件故障模拟方面也取得了不错的成绩。近些年, 国内的其他高校和科研部门也开始重视故障注入的研究工作, 如清华大学、北京航空航天大学、航天部 502 所、航天部 771 所等单位也开始开展这方面的研究工作。

国内对虚拟机故障注入的研究不是很多, 浙江大学的车建华等人基于软件模拟的方法针对 VMM 的管理功能做了很多的故障注入工具^[43], 涉及的故障类型有虚拟机状态管理功能故障、虚拟机资源调度策略故障、虚拟机迁移故障、虚拟机性能隔离故障、虚拟机安全隔离故障等。文献 [30-44] 深入研究了 Xen 的实现细节。总结了实时迁移的相关技术, 特别是 Xen 动态迁移的原理, 并且结合源代码分析了 Xen 动态迁移的技术细节。曾凡平等提出的 XFISV (Xen-based fault injection technology for software vulnerability test) 就建立了一种有效且通用的软件测试模型, 和普通评测软件健壮性只从软件自身方考虑问题的方法不同, 此模型通过向软件和操作系统 (Xen 虚拟机) 的交互层注入故障, 观测软件的反应来评价软件的健壮性。南京大学的王飞飞等人开发的 RandHyp 工具, 用来保护 Xen 虚拟机的超级调用 (Hypercall) 机制, 他们利用随机技术, 在 Xen 内部构建一种透明的机制, 能够对有威胁的超级调用进行阻止, 保护系统不受威胁。此工具虽然不是用来做故障注入的, 但其对超级调用的研究对于设计基于超级调用的故障注入工具有很大的帮助。

近几年, 在国家高技术研究发展计划 (863 计划) 重大项目的资助下, 哈尔滨工业大学容错计算团队在研究云平台虚拟化技术的基础上, 针对虚拟机的各个层次, 开发了相应六种故障注入工具, 并针对虚拟机的关键技术设计出了特定的故障注入方法^[45-48]。

如图 4 所示, 这六种故障注入工具均由软件的方式实现, 其中面向虚拟机监控器 Hypervisor 的有: 超级调用故障注入工具, 事件通道故障注入工具, 内存故障注入工具。面向虚拟机的有: 内核内存故障注入工具, 寄存器故障注入工具 (这两个针对 DomU), 管理功能故障注入工具 (针对 Dom0)。下面逐一进行简单的介绍。

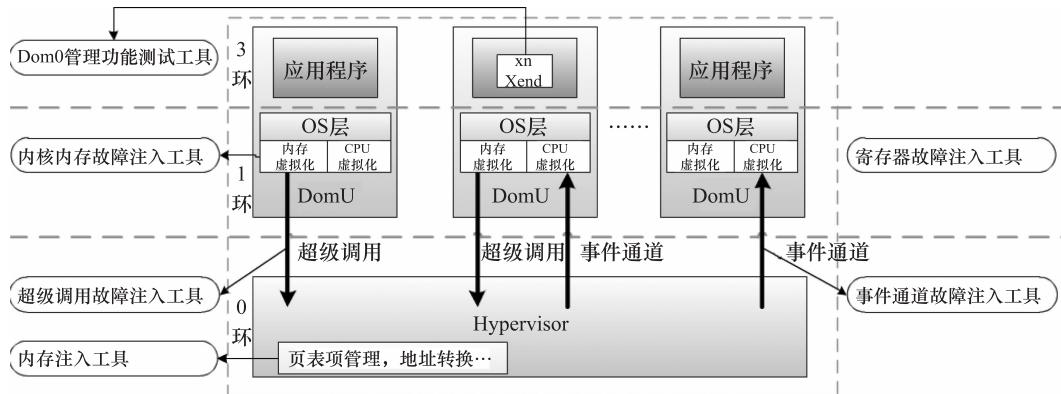


图 4 XEN 虚拟化环境的六种故障注入工具

1) VMM 超级调用故障注入工具。此工具针对 XEN 半虚拟化所依赖的重要机制——超级调用而设计实现，在 Hypervisor 中拦截超级调用的调用号及参数，然后进行修改并观察系统虚拟机 VM 和 VMM 的反应。在故障注入过程中，用户态可以通过程序控制故障注入的类型和持续时间。注入故障后，可以通过人为的虚拟机操作为系统增加负载。

2) VMM 事件通道故障注入工具。此工具针对 XEN 半虚拟化所依赖的时间通道机制而设计实现，XEN 中事件通道有一部分也是通过超级调用来完成，所以可以用给超级调用注入故障的方式来实现。还有一部分事件通道有自己的驱动处理程序，针对这部分也研究和设计了独立的方法来完成故障注入。

3) VMM 内存故障注入工具。此工具也是针对 XEN 的非常重要的特性——内存管理来实现的故障注入工具，我们知道对于 VMM 来说内存管理有个最基本的要求，即提供隔离性，不能让其中一个 VM 访问另外一个 VM 的内存，这样才能在安全和性能方面保证 VM 的独立性。此工具也是从这点出发，针对 VMM 的页表机制，通过修改页表项的值来实现访问不该访问的内存，从而达到故障注入目的。通过观察故障注入后 VMM 以及 VM 的反应来评测系统对于内存故障的容错性能^[44]。

4) DomU 内核内存故障注入工具。此工具在客户虚拟机（Guest OS）中实现，针对虚拟机的一项重要功能，即 Hypervisor 必须保证虚拟机之间的故障隔离性，不能因为一个 DomU 的失效而导致故障传播，致使其他虚拟机发生故障。在实现上，此工具在客户虚拟机内核中修改内核函数，致使内核运行异常并发生严重错误，导致系统宕机。注入故障后，通过观察故障虚拟机和其他虚拟机的运行情况以及 VMM 的运行情况，来评测 XEN 对于此类故障的容错能力。

5) DomU 寄存器故障注入工具。此工具在客户虚拟机（Guest OS）中实现，对于 XEN 以及所有虚拟机 Hypervisor 来说，给 VM 提供的硬件虚拟化资源中，最重要的即是内存和 CPU 的虚拟化，上面提到的 DomU 内核内存故障注入工具是对内存虚拟化隔离的验证，而寄存器故障注入工具是对 CPU 虚拟化隔离的验证。在实现上，通过在 Guest OS 内核态更改虚拟机寄存器的值及状态，达到故障注入的目的。通过观察故障虚拟机和其他虚拟机以及 Hypervisor 的运行状况，来评测 XEN 对于此类故障的容错能力。

6) Dom0 管理功能测试工具。Dom0 在 XEN 半虚拟化环境中发挥着至关重要的作用，其上运行着大多数实际物理硬件的驱动程序。而且 Dom0 中运行着 XEN 管理其他虚拟机的进程，它提供控制其他 DomU 的接口，对其他虚拟机的创建、关闭、迁移、备份等操作都需要此接口来完成。此工具即是针对此功能而设计的测试方法。具体来说，可以在虚拟机进行密集的内存换页、页表更新、进程创建等的情况下通过不断修改多台虚拟机的状态实现对状态管理功能容错性测试；通过虚拟机在压力负载很高的情况下进行多台虚拟机在多台主机之间高频率的动态迁移，来实现虚拟机迁移功能的测试；通过在一个 Hypervisor 上的一个虚拟机中运行资源高消耗的程序（如 CPU 高消耗程序、内存耗尽程序），然后观察其他虚拟机对资源的需求是否受到影响来评测虚拟机性能隔离功能等。总之 Dom0 管理功能的健壮性对整个虚拟化平台的稳定性有至关重要的作用，上面三项是虚拟机 Hypervisor 提供的最重要管理功能，通过对这些项的测试达到对 Dom0 管理功

能的容错性能测试。

以上六个测试工具分布于 XEN 虚拟化环境的各个部分。VMM 超级调用故障注入工具和事件通道故障注入工具以及内存故障注入工具用来评测 Hypervisor 层的容错性能；VM 内核内存故障注入工具和寄存器故障注入工具用来评测客户虚拟机 DomU 的容错性能；VM 管理功能测试工具用来评测 Dom0 管理功能的健壮性。这些工具全面覆盖了 XEN 虚拟化结构的三个组成部分，有效模拟了实际运行过程中可能遇到的各种故障。

3.3 构件软件可靠性过程国内研究进展

从云计算软件开发技术视角来看，面向对象技术的出现与快速发展以及基于 Web 开发的广泛应用使得模块化、组件化软件开发模式进展加速，由此衍生出基于构件的软件 CBS (Component-Based Software) 及开发方法^[49]。构件软件已成为一种主流软件形态，广泛应用在各种云计算平台中，其可靠性问题受到了极大关注。

构件软件可靠性过程是指在软件整个生命周期内以可靠性为核心，不断提高可靠性的增长和发布实现预期目标的可靠软件的动态过程。针对构件软件的可靠性过程内容和实现技术，文献 [50] 中把描述集成测试中故障检测与改正的动态过程称为可靠性过程分析，文献 [51] 则侧重以仿真的方法来分析排错过程，而文献 [52] 基于马尔可夫模型进行可靠性测评。文献 [53] 对基于 NHPP (Non-Homogeneous Poisson Process) 类型的构件软件的可靠性增长模型进行了研究。文献 [54] 提出了一种改进的可加模型，首先获得构件的实际测试数据，并利用仿真将测试剖面转换为运行剖面，进而建立实现单一构件的 G-O 模型。

构件软件可靠性过程主要关注在构件软件的整个生命周期内，主要以动态提高可靠性这一过程为核心，实现预期可靠性相关目标。其研究的要点可以概括为：

- 1) 可靠性的建模表示、度量计算、预测、保证与评估技术。
- 2) 采用适当的技术手段实现对测试与运行过程进行动态描述，实现可靠性的动态增长。
- 3) 合理分配受限的测试资源，实现预期目标下的软件最优发布等。

可以看出，相比于软件可靠性工程中对可靠性的设计与管理、定量测评与保证技术，构件软件可靠性过程在其基础上所涵盖的技术内涵更为丰富，突出面向可靠性提高的全过程，是可靠性工程在考虑构件软件基本特征基础上面向可靠性动态提高研究的进一步发展。例如，以测试环节为例，以往可靠性工程中测试的目的是发现 Bug，以检测与修复为导向，而可靠性过程则以不断提高测试过程中的可靠性为目标。此外，可靠性工程显示出明显的软件工程特征，目前可靠性过程则并不显著。可靠性过程技术均是紧密围绕着构件软件的可靠性提高方面来展开，但对于应用在构件软件的某个生命周期、研究内容的差异性以及具体的技术形式，尚需要明确如何确定每类技术问题的核心，以及如何界定每类技术的边界。

如图 5 所示，构件软件可靠性过程技术主要是由构件软件的生命周期进程划分中涉

及的可靠性研究内容和技术上的总体范畴来决定的^[55]。涉及的关键技术主要有：

- 1) 可靠性的建模技术，描述如何用合适的模型建立构件的失效以及整个软件的可靠性求解思路。
- 2) 可靠性的测评技术，实现对整个构件软件可靠性的度量、预测。
- 3) 可靠性增长模型技术，基于测试或运行数据，采用统计分析的方法建立可靠性不断提高的数学模型。
- 4) 可靠性过程仿真技术，采用离散事件的仿真方法对构件软件的测试过程进行建模描述，进而建立一段时间内检测与排除的故障数量与可靠性的定量关系。
- 5) 测试资源分配与最优发布技术，建立测试过程中测试资源在构件间分配的最优化关系式，实现在预期目标下的最优发布软件。
- 6) 基于 PCM 的可靠性相关技术则是将模型驱动建模方法与传统可靠性研究相结合而产生的新技术，是可靠性研究与软件工程相互融合的新发展。

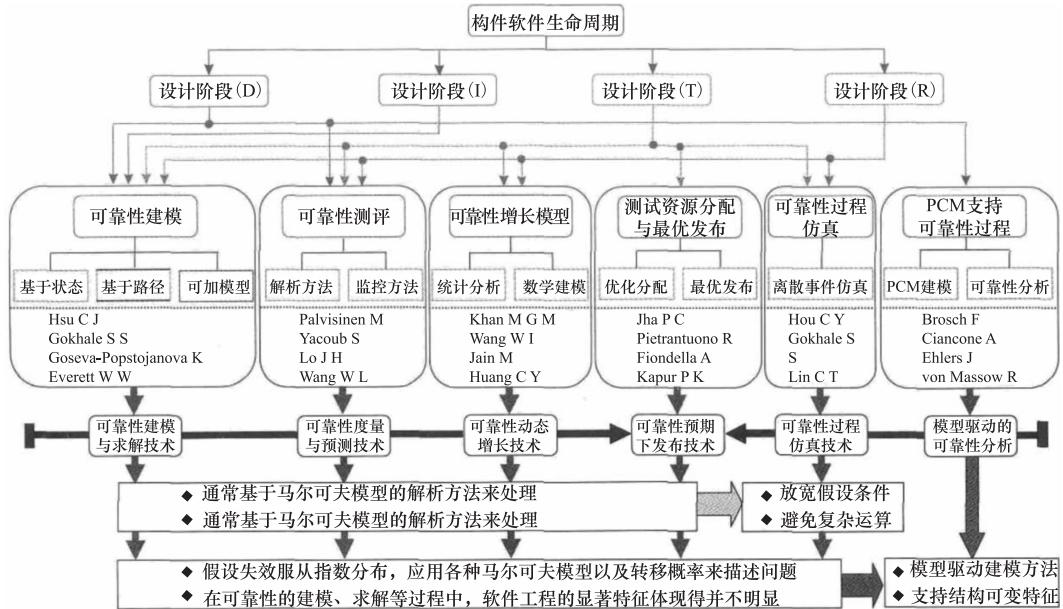


图 5 构件软件可靠性过程关键技术

基于体系结构的建模表示是构件软件可靠性过程研究的基础，其对构件软件的内部结构特征、构件失效率等进行了必要描述；可靠性增长模型则是可靠性过程研究的核心，其他研究问题的描述与求解过程中大都需要依托可靠性增长模型的定量评测能力来实施。

3.4 基于 BDD 的互联网络可靠性国内研究进展

1999 年，Kuo 提出的基于 OBDD 的网络可靠度算法 EED-BFS (Algorithm Based on Edge Expansion Diagram-Breadth-First-Search Ordering) 大大推动了网络可靠性评估算法的发展^[56]。该算法选择宽度优先搜索 (Breadth-First-Search, BFS) 策略进行变量排序，在

利用边扩张（Edge expansion, EP）构建网络 OBDD 的过程中，能有效识别同构子图，消除冗余节点，从而避免了大量的冗余运算，为大型网络可靠度的计算奠定了基础。但是其中并未介绍同构子网的识别方法，很难对其算法进行评估。

Hardy 提出了一种评估全网可靠度的方法^[57]。该方法利用分区等价类来识别网络中的同构子网，该方法比文献 [58] 中的方法更有效且能计算部分规模较大网络的可靠度，然而文献 [56-57] 中的算法存在两个不足：当网络规模很大时，OBDD 节点的数目增多，内存消耗急剧增加，算法效率下降；另外计算网络可靠度时需要分两步进行，首先生成网络的 OBDD，然后通过遍历 OBDD 计算网络的可靠度，计算过程较复杂。针对这些问题，文献 [58] 提出了一种改进方法，在 OBDD 的每个节点存储概率信息，每个 OBDD 节点只处理一次，之后可以释放，这使得算法只保存不多于两层的 OBDD 节点信息，有效节省了内存。最重要的是这种方法可以直接计算出网络的可靠度和失效率，降低了算法复杂度，提高了计算效率，是目前比较有效的计算网络可靠度的方法。

基于决策图的网络可靠性评估是一个很有前途的领域，但由于研究时间较短，相关的理论和实践都远不成熟，浙江师范大学网络可靠性研究团队在这方面的系统地研究了以下问题。

（1）网络可靠性 BDD 分析中边排序问题

排序问题在故障树分析的 BDD 方法中已经广为研究了。我们以故障树结构特征的核心要素“重复变量”为基础，从无重复故障树和带重复故障树两个不同的角度出发，研究了启发式变量排序方法^[59]。以深度优先遍历排序策略基础，通过不同权值机制设计多种具有基本事件编号无关的新型排序策略，并研究了策略性能和故障树结构特征的依赖关系^[60]。

类似地，网络可靠度 BDD 分析过程包含边排序、路径函数生成、BDD 生成和评估三个步骤，其中 BDD 生成和可靠度评估的计算复杂度与 BDD 尺度线性相关，而 BDD 尺度取决于边排序。因此排序问题是 BDD 分析方法研究的核心问题。在网络可靠性 BDD 分析中适当地选择边排序，能够有效缓解 BDD 尺度组合爆炸问题，加速系统的可靠性评估。导致最小尺度 BDD 的边排序称为最小边排序。优化边排序成为最小边排序是 NP 完全问题。通常采用一些启发性方法来生成近似最小的变量排序。为此需要研究网络的结构特征参数和可靠性评估性能指标参数；需要引入节点权值机制，设计多种与网络节点编号无关的新型排序策略；需要研究策略性能和网络结构特征的依赖关系，从而便于在实际系统分析时采用合适的分析方法^[61-64]。

另一方面，以往对边排序策略的性能研究仅仅侧重于平均性能分析层面，没有考虑排序起点对排序结果的影响，即相同策略不同排序起点被认为性能相同，也就是没有很好地回答“策略 H 排序起点在网络源点 S 处是否性能最佳”这一问题。为此需要针对各种不同类型的网络，研究是否以源点 S 为排序起点的变量（边）排序是高性能排序、对于高性能变量（边）排序的起点是否和 ST 相关、高性能变量（边）排序的起点分布是否有特定规律等问题^[65]。

(2) 网络可靠性 BDD 生成算法性能改进问题

网络可靠度 BDD 分析中路径函数的生成过程中需要进行边扩展，如果不采用有效的优化技术就会产生许多无效扩展路径，形成严重的性能问题，因此避免无效路径扩展显得异常重要。为此，需要研究无效扩展路径的产生原因及其分类；需要利用图论算法的最新研究进展，根据不同的原因，设计相应的高性能消除技术^[66-67]。

另一方面，网络可靠度 BDD 分析中路径函数的生成过程可能产生许多同构子图，导致大量冗余计算，因此快速判断同构子图显得异常重要。为此，需要研究同构判定所需信息集合，包括点的信息集合和边的信息集合；需要研究信息集合的最小化定义；需要研究基于不同信息集合的判定性能和网络结构特征的依赖关系，从而便于在实际系统分析时采用合适的判定方法^[68]。

(3) 大规模网络可靠性近似分析问题

基于截断 BDD 的近似分析技术在故障树分析中已经广为研究了。针对已有截断 BDD 算法在 BDD 操作和子 BDD 之间存在多对一的性能问题，文献 [69] 基于回写机制设计了改进截断 BDD 算法，结合 0 截断和 1 截断，利用迭代逼近的策略，设计了静态故障树模型的可靠度区间近似分析方法。针对已有故障树基准测试库设计中存在的问题，文献 [70] 提出了自动生成结构特征多样化的随机故障树生成算法，并应用于变量排序策略的性能比较。

类似的，相对于传统的最小路集和最小割集方法，采用精确的 BDD 算法确实能够有效扩大可分析网络的规模，但是由于网络可靠性评估问题本身固有的复杂性，当网络规模达到一定限度（比如 12×12 ）时，已有的精确算法不能获得有效的结果。采用截断 BDD 对布尔表达式进行近似分析是当前可靠性工程领域的研究热点，但是针对网络可靠性的截断 BDD 可靠性评估研究很少。为此，需要研究基于截断 BDD 的大规模网络可靠性近似分析方法，需要研究基于路径函数截断的网络可靠度近似分析，包括 1 截断的上界分析和 0 截断的下界分析，以下截断误差分析、截断路集分析等。

值得一提的是，国内关于网络可靠性评估的研究稍晚于国外，但是跟踪和发展非常迅速。截至目前，北京航空航天大学可靠性与系统工程学院、北京大学无线电系、浙江师范大学、北京邮电大学经济管理学院、国防科技大学信息系统与管理学院、总参 61 所等多家单位都有网络可靠性研究机构。其中，北大主要从通信领域的角度研究网络可靠性分析、评估技术；北邮主要从管理角度研究通信网可靠性的管理与综合评价方法；总参 61 所主要从物理学角度研究基于云模型的计算机网络可靠性评价方法；国防科大主要从复杂性科学与系统工程的角度对复杂网络抗毁性与可靠性进行研究；北航主要从可靠性工程的角度研究网络可靠性共性评价、分析技术。

3.5 互联网络可靠性参数及优化国内研究进展

网络可以模型化为一个图，因此网络的很多性质可以通过研究图的各种参数来达到。

在通信网络的设计与维护过程中，网络可靠性是一项重要的性能指标，其主旨就是在网络的部分元件发生故障的情况下力求使网络的破坏程度尽可能小。

当网络承载越来越多的服务和应用时，网络对各种性能的容错能力（fault-tolerance）就成为网络设计中需要考虑的重要因素，即最多多少部件发生故障时剩余网络仍具有给定的性质。超点（边）连通性的容错度是一个较新的概念，尚存在大量有待深入研究的课题。

近年来，在国家优秀青年自然基金的资助下，浙江师范大学网络可靠性研究团队在网络可靠性参数及优化边的连通性方面得到如下成果。

（1）边的连通性

为了更精确地度量网络可靠性，Esfahanian 提出了限制性边连通度的概念，这一概念被 Fiol 等人推广到了 k -限制性连通性上。对于一个连通图 G 中的边子集 S ，如果 $G-S$ 不再连通，且 $G-S$ 的每个连通分支都至少含有 k 个点，则称为 k -限制性边割。最小 k -限制性边割中所含的边数称为 k -限制性边连通度，记作 λk 。与 λk 密切相关的一个概念是 k -阶等周边连通度 γk ，它是满足如下要求的最小边割中所含的边数：该边割把一个连通图分成两部分，每部分至少有 k 个点。它与 λk 的区别就在于不要求这两部分导出的子图都连通。正因为这一区别， γk 对于任意 $k \leq |V(G)|/2$ 都有定义，而限制性边连通度 λk 的存在性则复杂得多。

在文献 [71] 中，我们首次揭示了 γk 与 λk 的关系。对于一个 k -正则图，当 k 不大于围长的二倍时有 $\gamma k = \lambda k$ 。这一结果使 λk 的研究在 k 不大于 6 时都得以简化，给出了 γk 的一个上界 βk （它是 k 个点的邻边数的最小值），而 λk 的上界 ξk 是我们在前期工作中给出的（ βk 与 ξk 的不同之处在于后者要求 k 个点的导出子图连通，且剩余 $n-k$ 个点的导出子图也连通）。我们把达到相应上界的图分别称为 γk -最优图和 λk -最优图，并研究了这两个参数达到最优的充分条件：①除了完全二部图 $K_3, 3$ 和完全图 K_4 ，任一 3 -正则点传递图或边传递图是 γk -最优图的充要条件是其围长至少为 $k+2$ 。②任一 d -正则的边传递图 G ，只要 $d \geq 6ek(G)/k$ ，那么 G 就是 γk -最优图，其中 $ek(G)$ 为 G 中 k 点导出子图的最大边数？

（2） Rk -边连通度与圈边连通度

对于连通图 G 的边割 S ，如果 $G-S$ 中的每个点都至少有 k 个好的邻点，则称为 Rk -边割。最小 Rk -边割中所含的边数记作 λk 。该定义在 $k=1$ 时与 $\lambda 1$ 吻合，但在 $k \geq 2$ 时其研究难度远大于 λk 。事实上，我们证明了当最小度 ≥ 3 时， $\lambda 2$ 与圈边连通度 λc 相等。圈边连通度是在研究四色问题中提出的一个概念，在很多经典图论问题的研究中都具有很重要的作用。以往的研究大多是利用圈边连通度的值作为研究其他问题的条件，对该参数本身的研究并不多。从上述结果来看，它在衡量网络可靠性方面也具有一定作用，因此研究它的最优性是很有必要的。

我们研究了对称图的 Rk -边连通度与圈边连通度，得到如下成果。在文献 [72] 中，我们给出了 λc 的一个普适性的上界 ζ ，它是最短圈的最小关联边数。我们将 $\lambda c = \zeta$ 的图称

为最优圈边连通图，并研究了传递图中圈边连通度的最优性，证明了所有最小度 $\delta \geq 4$ 的边传递图都是圈边最优图。所有正则度 $k \geq 4$ 、围长 $g \geq 5$ 的点传递图都是圈边最优图。

(3) 点连通度、超点连通性和高阶限制性点连通度

对于限制性点连通度，因其研究难度远大于边的情形，所以目前国际上相关结果还比较少。在前期工作中，我们研究了传递图的 super- κ 性和 hyper- κ 性，并首次提出了 semi-hyper- κ 的概念。如果一个图的任一最小点割都是某一最小度点的邻点集合，则称该图是 super- κ 的。Boesch 等人证明了：作为网络拓扑结构，super- κ 图比非 super- κ 图更可靠。如果移去一个图的任一最小点割恰产生两个连通分支，且其中之一为单点，则这个图是 hyper- κ 的。显然，hyper- κ 图必为 super- κ 图，且这样的图在移去一个最小点割后的损坏程度小于非 hyper- κ 图。我们在前期工作中把 hyper- κ 的概念推广到了 semi-hyper- κ ，要求移去一个图的任一最小点割都恰好产生两个连通分支，并完整地刻画了 semi-hyper- κ 的边传递图。对于高阶限制性点连通度的研究，现有结果更少，我们在该方面也做了一些有益的尝试。

在文献 [73] 中，我们研究了 semi-hyper- κ 的点传递图，首次提出了块可约的概念，用它给出了一个点传递图是 semi-hyper- κ 的充要条件。对于正则度不大于 7 的点传递图，我们还找出了所有非 semi-hyper- κ 图。特别地，用生成元完整地刻划了非 semi-hyper- κ 的极小 Cayley 图。在文献 [74] 中，我们研究了边传递图的超点连通性，证明了除了圈 C_n 和立方体图 Q_3 的线图 $L(Q_3)$ 以外，所有不可约的边传递图都是超点连通图。在文献 [75] 中，我们又进一步挖掘了二部图中的原子无交性，刻画了点传递二部图的 super- κ 性和 hyper- κ 性^[76]。

(4) 网络数据副本最优放置问题

为了增强网络数据的可靠性，一个有效的方法是保持网络一定的冗余度，即在网络若干节点保存同一数据（称为数据副本），从而在网络的某些部件发生故障时仍能保证数据的可获取性。在哪些节点保存数据副本是该应用所关心的最主要的问题。该问题的最优解与网络拓扑结构有关，也与网络协议有关。在该方面，以下问题已完全解决：树状网络中各种协议下的最优解；环状网络中 read only、write only 和 read any/write all 等协议下的最优解；线性总线网络中 read any/write all 和 majority voting 协议下的最优解。在 [77] 中，我们刻画了具有偶数个节点的环状网络在 majority voting 协议下副本的最优放置方案，从而完全解决了这一个八年都未见进展的难题。

4 国内外研究进展比较

4.1 计算平台和互联网络可靠性评估研究进展比较

国外在基础计算系统相关的计算平台和互联网络的可靠性评估研究中，多采用状态

空间分析、组合分析和分治分析方法。其中，基于 CTMC 的状态空间方法只能处理指数分布失效，同时存在严重的状态爆炸问题；同样的组合分析方法即使对于一个简单的系统，组合模型的建立和求解也是非常繁琐的，有时甚至由于运算量巨大而无法求解；而分治分析方法利用全概率公式把原问题分解为不完全覆盖故障发生和不发生这两个互斥子问题，然后进行子问题求解和结果综合，对于分析技术人员要求较高，分析自动化方面存在缺陷。

国内近几年提出了在 MDD 框架下多故障模式计算平台统一建模方法，以及基于单一系统级 MDD 模型的多故障模式统一分析方法，从而简化分析模型和分析方法的复杂性，扩大可分析系统的规模和复杂度、提升可靠性评估的准确性和效率。

此外，国内近几年在基于 BDD 的互联网络可靠性评估方面也做了大量工作，针对网络可靠性 BDD 分析中边排序问题、网络可靠性 BDD 生成算法性能改进问题、大规模网络可靠性近似分析问题进行了深入研究，提出了一系列高效的算法和策略，大大推动了网络可靠性 BDD 方法的分析能力。

4.2 基于故障注入的基础计算系统可靠性评测研究进展比较

近年来，国内通过研究 Xen 体系结构，设计和开发了六种故障注入工具，可以对 Xen 进行全方位的测试。面向 Xen 底层 Hypervisor 的有超级调用故障注入工具、事件通道故障注入工具和内存管理故障注入工具，面向虚拟机 DomU 的有内核内存故障注入工具和 CPU 寄存器故障注入工具，面向虚拟机 Dom0 的有虚拟机管理功能测试工具。这些工具除了分布于 Xen 体系结构的各个方面，还主要针对 Xen 半虚拟化的核心技术超级调用和事件通道，以及虚拟化中最重要的部分——内存虚拟化和 CPU 虚拟化，还特别面向 Xen 中 Dom0 的特殊地位开发了相应的测试工具。

和国外的故障注入工具相比，国内所实现的故障注入工具没有采用修改原有代码的方式，而是采用更灵活的动态可加载内核模块，但国外所采用的内存故障注入方法对我们实现的 VMM 内存故障注入工具有很大帮助。此外，国外相关研究所提出的用户态的检查点库，对我们设计 VMM 事件通道故障注入工具有很大启发作用。

4.3 构件软件可靠性过程研究进展比较

从云计算软件开发技术视角来看，构件软件已成为一种主流软件形态，广泛应用于各种云计算平台中，其可靠性问题受到了极大关注。国外的构件软件可靠性过程研究，侧重以仿真的方法来分析排错过程或者基于马尔可夫模型进行可靠性测评。

而国内主要对基于 NHPP (Non-Homogeneous Poisson Process) 类型的构件软件的可靠性增长模型进行了研究，并提出了一种改进的可加模型：首先获得构件的实际测试数据，并利用仿真将测试剖面转换为运行剖面，进而建立实现单一构件的 G-O 模型。

4.4 互联网络可靠性参数的可实现性问题、性质和优化问题研究进展比较

因为网络的拓扑结构通常被模型化为图，所以图论中的一些参数可用于衡量网络的某些性能。国外网络可靠性参数研究集中在图的连通度 κ 和边连通度 λ ，一般说来，它们的值越大网络就越可靠。但是，点连通度和边连通度在刻画网络可靠性方面有着明显的不足，它们只能度量在最坏情况下网络的破坏程度，而低估了网络的弹性和容错能力。

针对这些不足，国内提出了各种各样的新的网络可靠性衡量参数，如①边连通度、超边连通性、限制性边连通度等周边连通度。② Rk —边连通度与圈边连通度。③点连通度、超点连通性、高阶限制性点连通度。④有向图中的限制性弧连通度。这些参数提升了网络可靠性的评估能力和准确性。

5 发展趋势和展望

5.1 多故障模式云计算平台可靠性研究展望

针对现代云计算平台故障模式越来越多、失效行为越来越复杂这一发展趋势，研究高效多故障模式云计算平台可靠性评估方法显得非常迫切，而且有着重要的现实意义。但是通过分析已有研究现状可知，相关可靠性评估方法研究还远未成熟，需要系统的解决以下问题。

1) 动态故障系统可靠性评估方面。在已有动态故障系统可靠性评估研究中，Dugan 等提出的模块化方法把动态门子树作为模块进行处理，当动态门子树较大时，所获得的模块过大，导致性能接近于传统的状态空间分析方法。另一方面，Chang 的模块化方法虽然对动态门子树可以进行模块化，但当动态故障树中出现较多的重复故障事件时，仍然会形成大模块，从而导致整体分析效率较低。为此，针对动态故障系统需要提出更细粒度的、能够充分利用静态结构特性的分析方法，从而简化分析模型和方法的复杂性，提升分析效率。

2) 共模故障系统可靠性评估方面。在已有共模故障系统可靠性评估研究中，针对一般性共模故障模型或者共模故障和动态故障共存的情况下，往往采用 SEA 方法或 EDA 方法对原问题进行分解，利用二值决策图分析技术进行子问题求解，基于全概率公式进行结果综合。该类分析方法对于具有少量共模故障的系统较为有效，当系统包含的共模故障数量增加时，子问题分解就存在组合爆炸问题，此外，显式构造子问题会使得整个分析过程较为复杂，难以全面自动化实现。为此，针对共模故障系统需要提出更高效的共模故障建模和分析方法，在统一框架下对各种类型共模故障和动态故障进行联合建模和

分析，从而简化分析模型和方法的复杂性，实现分析自动化，提升分析效率。

3) 传播故障系统可靠性评估方面。在已有传播故障系统可靠性评估研究中，EDA方法对于多传播故障共同作用的情况或者传播故障和其他类型故障共存的情况，存在模型复杂性和算法效率问题。而三值决策图方法只处理全局作用的不完全覆盖传播故障，忽略了选择性传播故障以及动态故障和传播故障共存的情况。为此，针对选择性传播故障、多传播故障、传播故障和其他类型故障共存等实际分布式计算系统，需要进一步扩展已有三值决策图方法，引入统一的建模框架，实现在单一框架下的全系统分析，提高可靠性建模和分析效率。

5.2 硬件组件维护策略研究展望

早期的维护策略研究主要面向二态单部件系统，随着系统复杂度的提高，以云计算平台为代表的现代复杂计算系统能够工作在多个性能等级下，基于系统可靠性的维护策略设计需要在大量部件状态分布的基础上对整个系统的状态分布展开计算，进而分析系统可靠性。这一工作采用多态系统模型对系统进行建模，并通过多态系统理论中面向多部件复杂系统的可靠性评价方法进行计算。基于多态系统模型的多部件复杂系统维护策略的研究成为目前维护策略研究的热点。然而多态系统大多结构复杂，由多种不同部件组成，各部件的状态对系统的影响都会很大。在维护操作选择较多的情况下，寻求优化策略和评价系统性能都比较复杂。

例如，可以采用等效年龄模型对不完美 PM 进行建模，借助通用生成函数（Universal Generating Function，UGF）评估系统可靠性，并采用遗传算法在大量的维护操作选择中对策略进行优化。该启发式方法在求解多态多部件系统的维护策略中取得了良好效果。实际系统中往往是在一定的费用或时间等限制下，要求系统在一个任务阶段满足一定的可靠性要求。在满足限制条件的情况下选择合适的维护操作使系统获得最佳性能的问题称为选择性维护问题。针对具有最小修和更换维护操作的情况，可以采用分支定界（Branch and Bound，BB）方法和启发式方法对部件失效率与年龄相关的系统优化选择性维护策略，以减少计算时间。结果显示启发式方法能够在较快的时间内获得接近最优解的策略。考虑到不完美 CM 和不完美 PM 是主要的维护操作，可将不完美维护模型整合到选择性维护策略中，并将维护操作离散化为可编码形式以便通过遗传算法求解，以保障到下一阶段任务结束时系统具有尽量高的可靠性。

综上所述，面向多态多部件的复杂系统维护策略优化的研究中主要存在以下几点问题。

1) 部件间存在相关性，包括经济相关性、结构相关性和随机相关性。经济相关性是指与单部件系统相比，对多部件系统进行分组维护操作时，由于规模经济的原因使平均费用降低，或者由于更高的停机时间使维护费用上升。随机相关性是指部件的状态影响到其他部件的失效时间分布。结构相关性是指多个部件在结构上形成一个整体，对一个失效部件的维护操作也意味着对工作部件的维护。

2) 部件维护需考虑对系统整体的影响。传统的维护策略优化中通常只考虑维护操作本身引起的费用。多态系统的维护中需要从系统角度对策略进行评价, 维护操作应该基于系统的状态和性能进行优化而不是单个部件的状态。在多部件系统中对单个部件进行维护, 当具有不同的 FRU 级别时, 可能会引起子系统不可用甚至整个系统停机。而维护的目标并不是使单个部件的状态更好, 而是使系统获得较好的整体性能。分析不同部件的维护操作对系统整体性能的不同影响, 采取适合的差异化维护策略是多部件系统维护中的重要问题。

3) 在维护资源受限制的情况下对维护操作进行选择。由于大型复杂系统整体造价较高, 受到维护费用和时间限制, 因此对系统中的全部部件采取效果最好的维护操作在很多情况下是不允许的, 同时又要求系统在一定任务期内保持较好的性能。如何在不同系统结构特征下于各类部件的大量维护操作中做出选择以获得更合理的维护计划, 这是多部件系统维护策略的研究难点。

5.3 互联网络可靠性评估研究展望

互联网络可靠性评估不仅可以在网络的规划阶段提供方案选择依据, 还可以在网络的运行验收阶段提供评价网络“好坏”的指标, 更可以在优化设计阶段提供指导。相应的评估方法是近年来一个活跃的研究领域, 但各类方法还存在各自的局限性。

1) 连通可靠性的评估方法是最早开始同时也是近年来研究成果最多的, 但对于大型复杂网络来说, 改进后的精确算法在提高计算效率上的作用并不大。因此, 近似算法的主要研究成果集中在仿真方法上, 尤其是对蒙特卡罗仿真方法的改进和创新应用。此外, 对于多态和共因故障的研究是近几年新兴的热点。因此, 在连通可靠性评估方面还需要对大型复杂网络或无线网络的仿真方法以及多态、共因故障的处理方法进行进一步研究。

2) 在容量可靠性评估方面, 随机流网络概念的提出是一个很大的进展, 但目前该领域的成果较少, 近 10 年的新方法主要是基于网络最小割集的方法, 其计算效率有待提高。

3) 性能可靠性是近几年的研究新热点, 但由于网络的复杂性、动态性、多态性等特点, 性能可靠性的评估难度非常大。近几年的相关研究围绕网络的时延、拥塞等故障因素, 但并没有一个综合统一的性能可靠性评估模型, 相关的完整评估指标体系也没有建立。

从以上 3 类网络可靠性评估方法的研究进展中可以看出, 网络可靠性的内涵在不断延伸, 关注用户需求及网络性能的评估方法逐渐成为研究热点。近几年来还出现了以业务为中心的可靠性评估方法, 它将“网络对某业务的支持能力”作为业务可靠性的度量。业务可靠性在前 3 类研究的基础上, 从用户的角度出发, 关注网络上运行业务对可靠性的影响, 从而反映网络对用户业务的满足能力, 是未来重要的研究趋势。在相关的可靠性参数体系、评估模型与算法、仿真和试验方法等领域已有一些初步的研究成果, 还需要针对网络中软硬件“耦合”故障、业务剖面描述和生成等问题进行更深入的研究。

6 结束语

应用于金融行业、电信行业、能源领域、交通行业、航空业等关系到国家经济安全和社会安全的关键领域的基础计算系统不仅要具有极强的事务处理能力，同时要具有很高的可用性，可长期提供高速、稳定、持续的信息处理服务。开展基础计算系统的可靠性研究，对于保障国民经济平稳运行具有重要意义。

本报告重点分析云计算平台中的多故障模式可靠性评估技术、故障注入技术、软件可靠性技术和硬件组件维护策略研究，以及互联网络中的网络可靠性评估技术和可靠性参数优化技术研究现状，综述了浙江师范大学可信计算团队和网络可靠性研究团队以及哈尔滨工业大学容错计算团队的研究成果，并对基础计算系统可靠性的下一步研究进行了展望。

参考文献

- [1] Michael Armbrust, Armando Fox, Rean Griffith, et al. A View of Cloud Computing [J]. Communications of the ACM, 2010, 53(4) : 50-58.
- [2] 罗军舟, 金嘉晖, 宋爱波, 等. 云计算: 体系架构与关键技术 [J]. 通信学报, 2011, 32(07) : 3- 21.
- [3] 钟伟彬, 周梁月, 潘军彪, 等. 云计算终端的现状和发展趋势 [J]. 电信科学, 2010, 26(3) : 22-25.
- [4] 莫毓昌, 杨孝宗, 崔刚, 等. 一般阶段任务系统的任务可靠性分析 [J]. 软件学报, 2007, 18(4) : 1068-1076.
- [5] 莫毓昌, 杨孝宗, 刘宏伟. Phased Mission System Reliability Analysis Based on Markov Regenerative Stochastic Process [J]. 宇航学报, 2006, 27(6) : 1335-1340.
- [6] J B Dugan, B Venkataraman, R Gulati. DIFtree a software package for the analysis of dynamic fault tree models [C]. Proceedings of the IEEE annual reliability and maintainability symposium, Philadelphia, PA, 1997 : 64-70.
- [7] KJ Sullivan, J B Dugan, D Coppit. The Galileo fault tree analysis tool [C]. Proceedings of the 29th annual international symposium on fault-tolerant computing, Madison, Wisconsin, 1999, 232-37.
- [8] C Y Huang, Y R Chang. An improved decomposition scheme for assessing the reliability of embedded systems by using dynamic fault trees [J]. Reliability Engineering and System Safety, 2007, 92(10) : 1403-1412.
- [9] Yuchang Mo, Quansheng Yang. Modular solution of dynamic multiple- phased systems [J]. Journal of Southeast University (English Edition), 2009, 25(3) : 316-319.
- [10] Y S Dai, M Xie, K L Poh, et al. A model for correlated failures in N-version programming [J]. IIE Transactions, 2004, 36(12) : 1183-1192.
- [11] Z Tang, H Xu, J B Dugan. Reliability analysis of phased mission systems with common cause failures [C]. Proceedings of Annual Reliability Maintainability Symposia, 2005, 313-318.

- [12] L Xing, L Meshkat, S K Donohue. Reliability analysis of hierarchical computer-based systems subject to common-cause failures [J]. Reliability Engineering and System Safety, 2007, 92(3) : 351-359.
- [13] L Xing, A Shrestha, L Meshkat, et al. Incorporating common-cause failures into the modular hierarchical systems analysis [J]. IEEE Transactions on Reliability, 2009, 58(1) : 10-19.
- [14] G Levitin, L Xing. Reliability and performance of multi- state systems with propagated failures having selective effect [J]. Reliability Engineering and System Safety, 2010, 95(6) : 655-661.
- [15] G Levitin, L Xing, H Ben- Haim, et al. Multi- state systems with selective propagated failures and imperfect individual and group protections [J]. Reliability Engineering and System Safety, 2011, 96 (12) : 1657-1666.
- [16] S V Amari, J B Dugan, R B Misra. A separable method for incorporating imperfect fault- coverage into combinatorial models [J]. IEEE Transactions on Reliability, 1999, 48(3) : 267-274.
- [17] L Xing, J B Dugan. Analysis of generalized phased mission system reliability , performance and sensitivity [J]. IEEE Transactions on Reliability, 2002, 51(2) 199-211.
- [18] Y Chang, S V Amari, S. Kuo. OBDD- based evaluation of reliability and importance measures for multistate systems subject to imperfect fault coverage [J]. IEEE Transactions Dependable and Secure Computing, 2005, 2(4) : 336-347.
- [19] A Myers, A Rauzy. Efficient reliability assessment of redundant systems subject to imperfect fault coverage using binary decision diagrams [J]. IEEE Transactions on Reliability, 2008, 57(2) : 336-348.
- [20] L Xing, Gregory Levitin, Chaonan Wang, et al. Reliability of Systems Subject to Failures with Dependent Propagation Effect [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2013, 43(2) : 277-290.
- [21] L Xing, Gregory Levitin. Combinatorial Algorithm for Reliability Analysis of Multi- State Systems with Propagated Failures and Failure Isolation Effect [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2011, 41(6) : 1156-1165.
- [22] M Le, A Gallagher, Y Tamir. Challenges and Opportunities with Fault Injection in Virtualized Systems [C]. In First International Workshop on Virtualization Performance: Analysis, Characterization, and Tools, Austin, Texas, USA, 2008, 78-96.
- [23] A L Goel, K Okumoto. Time- dependent error- detection rate model for software reliability and other performance measures [J]. IEEE Transactions on Reliability, 1979, 28(3) : 206-211.
- [24] S Yamada, M Ohba, S Osaki. S-shaped reliability growth modeling for software error detection [J]. IEEE Transactions on Reliability, 1983, 32 : 475-484.
- [25] Y Zhou, J Davis. Open source software reliability model: an empirical approach [C]. Proceedings of the Fifth Workshop on Open Source Software Engineering, New York: ACM, 2005, 1-6.
- [26] Xiang Li, Yanfu Li, Min Xie, et al. Reliability analysis and optimal version; updating for open source software [J]. Information & Software Technology, 2011, 53(9) : 929-936.
- [27] Y Tamura, S Yamada. Software reliability growth model based on stochastic differential equations for open source software [C]. Proceedings of the 4th IEEE international conference on mechatronics, Kumamoto, 2007, 1-5.
- [28] Y Tamura, S Yamada. Optimization analysis for reliability assessment based on stochastic differential equation modeling for open source software [J]. International Journal of Systems Science, 2009, 40(4) : 429-438.

- [29] 祁鑫, 左德承, 张展, 等. 多态多部件系统维护策略综述 [J]. 智能计算机与应用, 2013, 3(05) : 9-13.
- [30] S Frickenstein, L Whitaker. Age replacement policies in two time scales [J]. Naval Research Logistics, 2003, 50 : 592-613.
- [31] X Zhou, L Xia, L Jay. Opportunistic preventive maintenance scheduling for a multi-unit series system based on dynamic programming [J]. International Journal of Production Economics, 2009, 18 : 361-366.
- [32] 江逸楠, 李瑞莹, 黄宁, 等. 网络可靠性评估方法综述 [J]. 计算机科学, 2012, 39(5) : 9-13.
- [33] M Hayashi, T Abe. Evaluating Reliability of Telecommunications Networks Using Traffic Path Information [J]. IEEE Transactions on Reliability, 2008, 57(2) : 283-294.
- [34] G. Levitin. Reliability Evaluation for Acyclic Transmission Networks of Multi-state Elements with Delays [J]. IEEE Transactions on Reliability, 2003, 52(2) : 231-237.
- [35] H. Liu, M. Shooman. Reliability Computation of an IP/ATM Network with Congestion [C]. Proceeding of Annual Reliability and Maintainability Symposium, 2003 : 581-586.
- [36] H Liu, M Shooman. Simulation of Computer Network Reliability with Congestion [C]. Proceeding of Annual Reliability and Maintainability Symposium, 1999 : 208-213.
- [37] L Xing, Y Dai. A new decision diagram based method for efficient analysis on multi-state systems [J]. IEEE Transactions Dependable and Secure Computing, 2009, 6(3) : 161-174.
- [38] S V Amari, L Xing, A Shrestha, et al. Performability Analysis of Multi-State Computing Systems Using Multi-Valued Decision Diagrams [J]. IEEE Transactions on Computers, 2010, 59(10) : 1419-1433.
- [39] Yuchang Mo, Liudong Xing, Suprasad V Amari. A Multiple-Valued Decision Diagram Based Method for Efficient Reliability Analysis of Non-repairable Phased-Mission Systems [J]. IEEE Transactions on Reliability, 2014, 63(1) : 320-330.
- [40] Yuchang Mo, Liudong Xing, J B Dugan. MDD-Based Method for Efficient Analysis on Phased-Mission Systems with Multimode Failures [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2014, 44(6) : 757-769.
- [41] Yuchang Mo. A multiple-valued decision-diagram-based approach to solve dynamic fault trees [J]. IEEE Transactions on Reliability, March 2014, 63(1) : 81-93.
- [42] Yuchang Mo, Liudong Xing. An Enhanced Decision Diagram-Based Method for Common-Cause Failure Analysis [J]. Journal of Risk and Reliability, October 2013, 227(5) : 557-566.
- [43] 车建华, 何钦铭, 陈建海, 等. 基于软件模拟的虚拟机系统故障插入工具 [J]. 浙江大学学报(工学版), 2011, 45(4) : 614-620.
- [44] 江雪. 基于 Xen 虚拟机的动态迁移技术研究 [D]. 上海交通大学, 2009.
- [45] 刘伟娜. 面向安腾架构的高端容错机故障注入平台的设计与实现 [D]. 哈尔滨工业大学, 2010.
- [46] 王波. 高端容错计算机故障注入工具的设计与实现 [D]. 哈尔滨工业大学, 2011.
- [47] 秦磊. 面向安腾 2 处理器的故障注入工具设计与实现 [D]. 哈尔滨工业大学, 2010.
- [48] 冯刚. 面向云计算平台的虚拟机故障注入工具研究与设计 [D]. 哈尔滨工业大学, 2013.
- [49] 梅宏, 陈锋, 冯耀东, 等. ABC: 基于体系结构、面向构件的软件开发方法 [J]. 软件学报, 2003, 14(4) : 721-732.
- [50] 侯春燕, 崔刚, 刘宏伟. 基于率的构件软件可靠性过程仿真 [J]. 软件学报, 2011, 22(11) : 2749-2759.
- [51] S Gokhale, M Lyu. A simulation approach to structurebased software reliability analysis [J]. IEEE

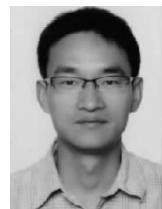
- Transactions on Software Engineering, 2005, 31(8) : 643-656.
- [52] SGokhale, W Wong, J Horgan, et al. An analytical approach to architecture-based software performance and reliability prediction [J]. Performance Evaluation, 2004, 58(4) : 391-412.
- [53] 侯春燕. 构件软件的 NHPP 类软件可靠性增长模型的研究 [D]. 哈尔滨工业大学, 2011.
- [54] 侯春燕, 崔刚, 刘宏伟. 实现构件软件可靠性分析的改进的可加模型 [J]. 高技术通讯, 2011, 21(3) : 267-272.
- [55] 张策, 崔刚, 刘宏伟, 等. 构件软件可靠性过程技术 [J]. 计算机学报, 2014, 37(12) : 2586-2612.
- [56] S Kuo, S Lu, F Yeh. Determining terminal-pair reliability based on edge expansion diagrams using OBDD [J]. IEEE Transactions on Reliability, 1999, 48(3) : 234-246.
- [57] G Hardy, C Lucet, N Limnios. K-terminal network reliability measures with binary decision diagrams [J]. IEEE Transactions on Reliability, 2007, 56(3) : 506-515.
- [58] J U Herrmann, S Soh. A memory efficient algorithm for network reliability [C]. Proceedings of the 15th Asia-Pacific Conference on Communications, Shanghai, China, 2009 : 703-707.
- [59] Yuchang Mo, Huawei Liu, Farong Zhong, et al. Depth-first event ordering in BDD-based fault tree analysis [J]. Computing and informatics, 2012, 31(6) : 1401-1416.
- [60] Yuchang Mo, Farong Zhong, Huawei Liu, et al. Efficient ordering heuristics in BDD-based fault tree analysis [J]. Wiley Journal of Quality and Reliability Engineering International, 2013, 29(3) : 307-315.
- [61] 潘竹生, 莫毓昌, 赵建民. 优先级边排序策略及其性能分析 [J]. 计算机科, 2014, 41(8) : 81-86.
- [62] 潘竹生, 莫毓昌, 钟发荣, 等. 一种新的启发式边排序策略及其性能分析 [J]. 计算机工程与科学, 2014, 36(11) : 2119-2129.
- [63] 伍欢, 钟发荣, 莫毓昌, 等. 网络可靠性分析中 BFS 策略与 POS 策略的性能比较 [J]. 山东大学学报 (工学版), 2015, 45(2) : 43-49.
- [64] Zhusheng Pan, Yuchang Mo, Liudong Xing, et al. New insights into breadth-first search edge ordering of regular networks for terminal-pair reliability analysis [J]. Journal of Risk and Reliability, March 2014, 228(1) : 83-92.
- [65] Yuchang Mo, Liudong Xing, Farong Zhong, et al. Choosing a heuristic and root node for edge ordering in BDD-based network reliability analysis [J]. Elsevier Journal of Reliability Engineering and System Safety, 2014, 131 : 83-93.
- [66] 刘轩, 潘竹生, 钟发荣, 等. 工程网络可靠性分析的网络简化方法 [J]. 山东大学学报 (工学版), 2015, 45(2) : 27-33.
- [67] 潘竹生, 莫毓昌, 钟发荣, 等. 网络可靠度 BDD 分析算法的性能改进 [J]. 计算机工程与科学, 2012, 34(9) : 26-33.
- [68] 陈荣根. 基于 BDD 的网络可靠性分析方法研究 [D]. 浙江师范大学, 2014.
- [69] Yuchang Mo, Farong Zhong, Xiangfu Zhao, et al. New results to BDD Truncation Method for Efficient top Event Probability Calculation [J]. Nuclear Engineering and Technology, 2012, 44(7) : 755-766.
- [70] Yuchang Mo, Jianmin Han, Zhizhen Zhang, et al. Approximate Reliability Evaluation of Large-scale Distributed Systems [J]. Journal of Information Science and Engineering, January 2014, 30 : 25-41.
- [71] Z Zhang. Extra edge connectivity and isoperimetric edge connectivity [J]. Discrete Mathematics, 2008, 308 : 4560-4569.
- [72] B Wang, Z Zhang. On cyclic edge-connectivity of transitive graphs [J]. Discrete Mathematics, 2009,

309 : 4555-4563.

- [73] Z Zhang. Semi-hyper-connected vertex transitive graphs [J]. Discrete Mathematics, 2009, 309 : 899-907.
- [74] Z Zhang, J X Meng. Super-connected edge transitive graphs [J]. Discrete Applied Mathematics, 2008, 156 : 1948-1953.
- [75] X D Liang, J X Meng, Z Zhang. Super connectivity and hyper connectivity of vertex transitive bipartite graphs [J]. Graphs and Combinatorics, 2007, 23 : 309-314.
- [76] Y F Zhu, Z Zhang. Restricted connectivity of line digraphs [J]. 数学研究, 2010, 43 : 107-113.
- [77] Z Zhang, W L Wu, S Shekhar. Optimal placements of replicas in ring network with majority voting protocol [J]. Journal of Parallel and Distributed Computing, 2009, 69 : 461-469.

作者简介

莫毓昌 浙江师范大学教授, 博士, 美国电气和电子工程师协会(IEEE)高级会员, CCF容错计算专委会委员, 主要研究方向为高可靠计算和复杂计算系统可靠性评估。



杨孝宗 哈尔滨工业大学教授, 博士生导师, CCF容错计算专委会资深会员, 主要研究方向为容错计算、可信系统与网络、移动计算。



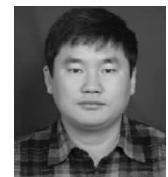
左德承 哈尔滨工业大学教授, 博士, 博士生导师, CCF容错专委会副主任, 主要研究方向为容错计算、可信系统与网络、移动计算。



张昭 浙江师范大学教授, 博士, 博士生导师, CCF会员, 国家自然基金优秀青年基金获得者, 主要研究方向为网络可靠性、组合优化算法、近似算法。



刘宏伟 哈尔滨工业大学教授，博士，博士生导师，CCF 容错专委会委员，主要研究方向为容错计算、软件可靠性。



钟发荣 浙江师范大学教授，博士，CCF 会员，主要研究方向为容错计算、可信系统与网络。



网络信息安全科技与应用发展综述

CCF 计算机安全专委会

摘要

随着社会的信息化程度越高，信息网络和各种信息系统的安全就愈加重要的背景下，国家的各种基础设施对信息系统的依赖程度越来越高，这使得网络信息安全问题正在变得更加突出和迫切。网络信息安全领域瞬息万变，一些大事件发生过后，很快就会被新的事件推送出人们的视野。因此，经过系统的整理和编辑，帮助广大网络信息安全工作者提高对网络信息安全事件的理解，并根据当时形势，总结、思考、消化、吸收有益的经验。

关键词：网络与信息安全，科技与应用，综述

Abstract

With the information of the society, the security of information network and all kinds of information system is more important, the country's infrastructure is more and more dependent on the information system, which makes the network information security is becoming more prominent and urgent. Network information security field is constantly changing, some big events, will soon be a new event to push out people's vision. Therefore, after the system of finishing and editing, to help the majority of network information security workers to improve the understanding of the network information security incidents, and according to the situation, summary, thinking, digestion, absorption of useful experience.

Keywords: Network and information security, Technology and Application, summarize

1 引言

2014 年，我国的基础信息网络运行总体平稳，但是安全形势不容乐观。据权威统计，2014 年我国境内网站的仿冒钓鱼点成倍增长，遭境外攻击、控制事件不断增加，共有 6 116 个境外 IP 地址承载了 93 136 个针对我国境内网站的仿冒页面，仿冒页面数量同比增长 2.1 倍；境内被篡改网站达 36 969 个，同比增长 53.8%；40 186 个网站被植入后门，其中位于美国的 4 761 个 IP 地址通过植入后门控制了我国境内 5 580 个网站，被侵入的网站数量远超其他国家。

2 国外网络信息安全发展现状

2.1 各国网络信息安全动态

2.1.1 美国

2014年，网络空间安全纵深发展，网络空间博弈日益加剧。面对这种严峻形势，美国不断加快网络治理步伐，制定并实施了一系列网络安全战略和政策。

(1) 完善网络信息安全立法

2014年，美国参议院和众议院都提出了很多关于网络安全方面的议案，这些议案的主旨在于保障美国家安全，强化美国在网络空间的监控能力。美国参议院：7月8日，通过了《网络安全信息共享法案（CISA）》（2014年版）。美国众议院：5月24日，通过了修改后的《美国自由法案》（S.1599），该法案试图对美国国家安全局（NSA）监控互联网的行为进行有限度的约束，但被参议院投票否决；7月28日，通过了三项法案，包括《国家网络安全和关键基础设施保护法案》（H.R.3696），《关键基础设施研究和发展进步法案》（H.R.2952），《国土安全部网络安全队伍建设法案》（H.R.3107）。参众两院通过并经总统签署的法案：12月18日，奥巴马总统签署了这些法案——《2014年联邦信息安全现代化法》（S.2521），更新了《2002年联邦信息安全管理法案》；《2014年国家网络安全保护法》（S.2519）；《网络安全人员评估法案》（H.R.2952）；《2013年边境巡逻员薪资改革法案》（S.1691）。

(2) 加强关键信息基础设施安全

2月12日，美国白宫正式推出一项私营部门可自愿加入《网络安全框架》的项目，旨在加强电力、运输和电信等“关键基础设施”部门的网络安全，通过该框架，美国私营企业和政府部门可以联手加强“关键基础设施”的安全性与适应性。

4月17日，美国证券交易委员会发布金融企业信息安全路线图，确保金融企业在探测和应对网络攻击方面做好准备。

6月12日，美空军网络司令部与美空军土木工程中心签署合作协议，从14个方面开展合作，强化美空军工控系统安全管理，保障关键信息基础设施安全稳定。

7月1日，美国国家标准与技术研究院宣布，将在2015年上半年制定并使用一套新型物联网安全框架，明确物联网相关定义，并提供基于此框架的网络安全系统。

8月11日，美国白宫宣布，美国政府已成立了一个全新的数字服务部门团队（US Digital Service, USDS），以负责美国医保等诸多政府网站的数据服务。

(3) 加快网络扩军步伐

3月4日，美国国防部颁布新版《四年防务评估报告》，该报告称：至2019年，美

国将建立多达 133 支网络战部队，包括 13 支“国家任务部队”、8 支“国家支持部队”、27 支“作战任务部队”、17 支“作战支持部队”、18 支“国家网络防御部队”、24 支“国家网络防御维护部队”和 26 支“作战指挥与国防部信息网络防御部队”。其中，40 支为攻击性网络部队。

4 月 8 日，美国西点军校称计划成立网络战争研究院“军队网络学院”，扩大军队网络人才规模，为国家网络政策制定及国防安全储备高级网络人才，未来三年将培养 75 名网络军事人才。

5 月 15 日，美国国防信息系统局（DISA）颁布《2014—2019 年战略计划》。该计划提出国防信息系统局将建立一支符合当前技术发展趋势的未来部队，重点关注网络的移动性、安全性以及云计算基础设施的互操作性。

6 月 19 日，一位国防高级官员表示，经过数年的筹划，国防部网络司令部终于开始进入了实战阶段。“国家任务部队”在得到指令后，能够封锁或对抗国外发起的网络攻击。

9 月 11 日，美国空军在马里兰州沃菲尔德的空军国民警卫队基地建立一个新的网络作战机构，下设一支网络作战大队和一支情报监视侦查中队，旨在为空军提供 24 小时不间断的全球情报信息和态势感知，该机构投入将达 1 000 万至 2 500 万美元。

10 月 1 日，美国海军组建了信息优势部队司令部，这是海军最新的职能司令部，合并了以前由各分散的信息优势司令部管理的使命、职能和任务。

（4）研发网络战武器

1 月 14 日，据报道，美国国家安全局和五角大楼网络司令部等情报机构已在全球近 10 万台电脑中植入“计算机网络攻击”软件，这些软件一方面能够对这些电脑进行实时监控，另一方面则能为美国发起大规模全球网络攻击修建一条“数字高速路”。

2 月 6 日，作为“X 计划”的组成部分，美国国防部先进研究项目局与知名军工企业雷神公司签署了价值 980 万美元的合同。雷神公司的研发工作主要致力于使国防部能够利用精确的评估结果，尽早、快速地衡量与执行网络任务。

3 月 11 日，美国空军发布意见征询书，为六种网络战武器系统的设计寻求相关支持，指挥、控制、通信、情报和网络项目执行办公室负责管理这六种武器系统。

（5）全面实施全球监控计划

1 月，英国《卫报》获悉的一份机密材料显示，美国国家安全局每天大约会截取全球近 2 亿条短信息，并利用这些信息获取诸如用户地理位置、联系人网络和信用卡等隐私数据。

2 月 25 日，斯诺登曝光美、英、澳、加、新五国间谍机构组成了“五只眼”情报联盟，其间谍行为中包括一个发布虚假信息、操纵网络言论、影响网络安全的“网络魔术师”任务。

3 月，据德国《明镜》周刊报道，美国国家安全局自 2009 年起入侵华为电子邮件服务器，成功获取华为内部资料及高管在内的员工电子邮件记录，包括华为总裁任正非、董事长孙亚芳邮件往来也被监控。

8月13日，斯诺登接受美国《连线》杂志采访时称，美国国家安全局正在秘密研发一项名为“怪物大脑”（MonsterMind）的网络安全工具，该工具不仅能识别、追踪和阻止潜在的网络攻击来源，还能在无人监督的情况下自动采取反制措施。

9月14日，美国国家安全局和英国政府通信总部（GCHQ）联合筹备名为“藏宝图”（Treasure Map）的网络监控项目，旨在绘制全球互联网动态地图，实现对全球所有电子设备的匹配和实时监控。

10月，美国“截听”网站援引斯诺登泄露的绝密文件称，美国国家安全局曾进行名为“鹰哨兵”（Sentry Eagle）的情报项目，该项目为“特殊受控信息”（ECI），只有国家安全局高层及政府个别高级官员才能接触。

2.1.2 欧盟

2014年，欧盟在网络信息安全方面开展了一系列战略和政策：欧盟将网络安全列入年度国防政策重点，完成首版智能交通系统（ITS）新标准，酝酿组建互联网监管同盟，通过《欧盟公共采购电子发票指令》和网络中立法律，组建网络联合搜查组织“J-CAT”，网络安全局发布欧洲智能电网安全认证报告等。

（1）欧洲网络与信息安全部：网络攻击向移动终端转移

1月26日，欧洲网络与信息安全部（ENISA）发布了《2013年网络威胁透视报告》。报告指出了2013年最严峻的网络安全威胁，并分析了数字化大环境下网络威胁呈现出的新趋势。报告指出，当前最严峻的网络安全威胁依次为路过式下载攻击、蠕虫和木马以及代码注入攻击。过去的一年，网络攻击及其利用的工具日趋复杂化；许多国家已具备入侵他国政府机构及私人电脑的能力，娴熟的网络间谍、攻击活动不局限于一小撮国家范围内；攻击对象移动化，过去几年针对个人电脑的攻击模式和工具转向移动终端；大数据和物联网成为两大新战场。

（2）欧盟呼吁以国际化方式管理互联网

2月12日，欧盟委员会呼吁，应该以国际化的方式管理互联网，改变当前“以美国为中心”的互联网管理现状。全球范围内的大规模监控和情报活动让人们对于互联网安全失去信心，欧盟提议“重绘全球互联网管理的地图”。

（3）欧盟将网络安全列入年度国防政策重点

2月12日，在欧洲议会的安全和国防委员会上，欧盟军事和国防官员称，确保网络安全成为2014年度欧盟最高国防政策的重点领域之一。欧盟表示，为了使各成员国“远离网络渗透”，欧盟正在展开网络安全技术研究和扩大交流协作。

（4）欧盟完成首版智能交通系统（ITS）新标准

2月18日，欧洲媒体消息称，欧盟标准化机构ETSI和CEN确认，已根据欧盟委员要求完成车辆信息互联基本标准的制订。该标准将确保不同企业生产的交通工具之间能相互沟通，并能与道路基础设施配套。

（5）欧盟酝酿组建互联网监管同盟

2月19日，英国《计算》周刊报道，欧洲委员会计划进行深入讨论，研究不同国家

法律和网络监管权限，以建立互联网管理同盟。鉴于网络上存在大量跨境交易活动，欧洲委员会有必要鉴别欧洲境内众多相互矛盾的网络立法规定。

2.1.3 英国

2014年，英国在网络信息安全方面的重要举措是加强基础设施和基础设施网络安全、加强网络教育和开展网络培训、采用多种措施应对网络威胁等。

(1) 加强基础设施和基础设施网络安全

5月12日，英国光纤宽带网络部署进展快速。英国电信宣布，截至今年3月，光纤宽带网络用户覆盖已经超过1900万家，占到全部家庭的2/3以上，比预期提前一年多完成了这一目标。

10月8日，据英国政府网站报道，英国政府正在改善供应链中的网络安全。从10月1日起，如果投标政府的合同涉及敏感和个人信息，以及某些技术产品和服务的提供，所有供应商必须遵守新的网络安全标准。

11月3日，《中国日报》报道，英国4个研究团队共享250万英镑（约2463万元人民币）政府基金，用于应对因日益增加的黑客攻击而引发关注的工业控制系统的安全问题。

(2) 加强网络教育和开展网络培训

3月13日，英国政府企业、创意与技能部表示，政府打算把网络安全教育扩大到11岁以上的中小学生，将向学校提供新的网络安全教材，资助学校招募网络安全教师。

8月4日，英国情报机构“英国政府通讯总部”授权六所英国大学提供训练未来网络安全专家的硕士文凭。

10月7日，据新华网报道，英国政府宣布开设免费网上培训课程，帮助英国企业提高防范网络攻击的能力。培训对象主要涉及律师和会计行业，培训内容主要包括如何防范和处理常见的网络威胁，如何保护数字信息等。

(3) 采用多种措施应对网络威胁

2月6日，英媒体报道，斯诺登泄露的最新情报显示，英情报机构“政府通信总部”曾以黑客的技术破坏黑客之间的通信——利用DDoS技术，迫使一个黑客组织使用的网络聊天室断线。

3月31日，英国内政部宣布正式组建国家计算机紧急应对小组，以协调应对针对计算机网络系统的攻击。该小组的主要职能是应对“国家级别”的互联网安全突发事件，并为政府、企业和学术机构提供相关互联网安全建议和风险提示，开展与其他国家的国际合作。

9月16日，据英国政府网站报道，在伦敦召开的美英全球赛博安全创新峰会上，英国商务部长文斯·凯布尔宣布将针对英国赛博企业应对赛博安全威胁的发展思路，投入400万英镑的比赛资金。

10月19日，据英国媒体报道，英国政府召集谷歌、微软、脸谱和推特等互联网企业，商讨进一步打击网络极端主义信息的举措，并要求互联网企业主动上交有助于警方

追踪极端主义分子的任何有关信息，同时进一步严格了删除极端主义信息的尺度。

2.1.4 德国

2014年，德国在网络信息安全方面的重要举措是提倡组建欧洲通信网络、出台《数字议程2014—2017》战略、应对网络监控举措等。

(1) 提倡组建欧洲通信网络

2月16日，德国总理默克尔在每周发布的博客中表示，她于19日与法国总统奥朗德会晤，商讨建立一个独立的“欧洲通信网络”，避免欧洲电子邮件和数据“绕道美国”，被美国国家安全局（NSA）监听。4月5日，据外媒报道，德国和法国倡议建立一个欧洲内部的通信网络，以避免电子邮件和其他电子数据通过美国境内传输。美国对这一倡议提出批评，称此举可能会违反国际贸易法。

(2) 出台《数字议程2014—2017》战略

8月21日，新华社报道，为了打造一个具有国际竞争力的“数字强国”，德国经济部、内政部等20日联合推出《数字议程2014—2017》，包括在变革中推动“网络普及”、“网络安全”及“数字经济发展”三个重要进程。

(3) 应对网络监控举措

4月13日，德国《明镜》周刊报道，德国航空航天中心多台计算机感染木马等病毒，几个月来，外国情报机构开始攻击德国航空航天中心，多台计算机已被渗透了间谍软件。6月4日，德国联邦最高检察官称，德国以“涉嫌从事特务及间谍活动”为由，对NSA窃听默克尔手机进行立案调查。7月10日，默克尔的发言人称，之所以有离境要求，是由于美国在德国的情报活动引发质疑，也是呼应德国检察官正在开展的调查。

2.1.5 俄罗斯

2014年，俄罗斯在网络和信息安全方面的重要举措包括：制定信息技术和电子政务发展方案，出台《俄罗斯联邦网络安全战略构想（草案）》，完善网络和信息安全法律，研发自主可控的信息技术等。

(1) 制定信息技术和电子政务发展方案

1月8日，俄罗斯批准《信息技术产业发展路线图》，提出到2018年将信息技术产业的从业人员数量翻一番；使信息技术产业的平均增长率超过GDP的平均增长率；信息技术产业的产值从84.4亿美元提升至140.6亿美元；通过普及推广信息技术提高劳动生产率等。8月14日，发布《组织国家电子政务基础设施管理进程》草案，提出发展电子政务基础设施的八个方向。

(2) 出台《俄罗斯联邦网络安全战略构想（草案）》

1月10日，公布《俄罗斯联邦网络安全战略构想（草案）》，阐述了俄罗斯即将出台网络安全战略的必要性和适时性，明确了网络安全战略的原则和方向，以及网络安全战略在国家法规体系中的地位。

(3) 完善网络和信息安全法律

3月21日，俄罗斯议员建议剥夺破坏或窃取政府网络资源黑客的人身自由10~15年，增大破坏国家网络应承担的责任。在3月16日克里米亚公投的前几天，黑客攻击了俄罗斯总统办公厅和俄罗斯外交部网站。议员们认为这是反国家罪行，必须受到应有的惩罚。4月25日，俄罗斯推出新法规，进一步规范网络言行，要求国内所有博客作者必须进行实名注册，并且公开电子邮件地址，博客作者需要确认所发布信息的真实性，不得披露他人的个人信息，禁止发布暴力或黄色内容等。5月5日，俄罗斯通过《博客作者法》，8月1日生效。6月30日，通过刑法修正案，将对借助互联网煽动极端主义活动的行为追究刑事责任。7月4日，批准新的互联网法案，规定凡收集俄罗斯公民信息的互联网企业必须将这些数据存储在俄罗斯境内的服务器上，并于2015年9月1日生效。8月8日，俄罗斯颁布法令，要求禁止匿名连接公共场所的Wi-Fi无线网络。10月13日，俄罗斯新法律规定，从2015年1月1日起俄罗斯公民的个人数据只能储存在俄境内，由于苹果产品的iCloud服务不符合这一规定，届时将在俄罗斯禁止售卖iPhone和iPad。10月15日，俄罗斯通过新法案，旨在限制外国资本对俄罗斯媒体的所有权，从2016年1月1日开始生效。11月15日，俄罗斯国家杜马通过了网络反盗版法修正案，进一步巩固网络作品著作权。

(4) 研发自主可控的信息技术

4月29日，为了防止俄罗斯国家银行系统受到西方可能的损害，考虑建立本国的银行交易网络。7月11日，俄罗斯总统普京宣布，俄罗斯准备参与国际信息安全系统的研发。

7月22日，由于美俄之间的政治形势时而会变得紧张，且美方刚宣布了对俄方的新一轮制裁，这使得俄罗斯考虑通过限制使用微软软件，并以此作为对美国经济的“反击”。俄罗斯政府正在起草一份新的法案，这将推动政府机构和国有企业向本地软硬件供应商倾斜。

8月14日，俄新社报道，俄罗斯计划建立空间通信系统，可覆盖全球并保证通信的安全性，预计将花费656亿卢布（约合人民币112.04亿元），以保护国家领导和军队通信的安全。

9月23日，俄罗斯提出，计划在出现战争等紧急情况时对国内互联网进行管理，以应对西方可能对其断网的情况。

11月16日，俄罗斯计划创建自己的“维基百科”，以确保市民可以获得更多有关自己国家的“翔实和可信的”信息。

11月21日，俄罗斯内务部长表示，俄罗斯将建立能够监控电信网络乃至整个信息领域的国家系统。

2.1.6 日本

2014年，日本在网络和信息安全方面的重要举措包括：优化调整网络安全机构，扩充网络战力量，设立“网络安全日”等。

(1) 优化调整网络安全机构

5月4日，日本决定设立“内阁信息通信安全官”一职，负责统一指挥网络安全与防御。5月14日，日本自民党和公明党就计划向国会提交的“网络安全基本法案”达成一致。法案规定了国家在网络攻击应对过程中的责任，并要求设立以内阁官房长官为首的“网络安全战略本部”。7月10日，日本决定在内阁设立“网络安全战略总部”，以应对日益增长的网络攻击。8月24日，日本警视厅与信息安全公司及咨询公司签订协议，成立分析网络犯罪作案手法等的研究机构。10月1日，东京警视厅成立网络搜查指导室以应对网络犯罪。10月30日，日本首相任用山本一太担任新设职务“网络战略顾问”。11月7日，日本表示，将成立打击网络犯罪中心，提高分析能力，以期迅速展开侦查以减少损失。

(2) 扩充网络战力量

1月18日，日本开展了应对网络攻击的史上最大规模的演习。3月26日，日本自卫队正式成立“网络防卫队”，编制在90人左右，将以每天24小时态势监视防卫省和自卫队的网络，应对网络攻击。4月11日，为保护自卫队的信息网络安全，日本防卫省将从2014年起实行3年计划，在网络空间上设立多个“监视据点”，提高网络预警能力。5月13日，日本表示，正在为其武装部队设计程序，以保护国内核设施不受来自网络空间的打击。6月16日，日本NEC“网络安全工厂”正式投入使用。7月15日，日本引进美国的网络防御系统，检测网络异常活动。9月8日，日本防卫省决定于2016年起引入诸如向自卫队网络植入病毒等模拟攻击训练，强化网络防御能力的实效性。

(3) 设立“网络安全日”

1月23日，据日本共同社报道，日本政府在官邸召开信息安全政策会议官房长官菅义伟任主席。为针对企业和个人开展有关打击黑客攻击措施的宣传警示工作，日本政府决定将2月3日定为第一个“网络安全日”。菅义伟在会议伊始就黑客攻击表示，这是“在国家安全和危机管理上越来越重要的问题。有必要进一步加强支撑信息安全措施的人才培养和体制完善工作”。内阁官房此前将2月定为“信息安全月”，配合新设的“网络安全日”，将在东京召集专家举行研讨会。政策会议上，内阁官房负责人汇报了打击黑客攻击保护政府信息安全的相关工作情况。

(4) 研发信息安全关键技术

2月4日，日本宣布开发出一种新系统，可在瞬间从社交网站等网络空间的投稿留言中检索出个人信息，并将这些涉及隐私的内容转换成特殊符号，用于防止家庭住址和姓名等个人信息被恶意使用。4月19日，日本总务省官员表示，政府将借助大数据技术预测自然灾害，把现有气象记录等数据与民众在互联网上发布的消息相结合，进而分析特定地区出现灾害的可能性，增强防灾信息的准确性。5月6日，日本富士通公司成功开发出防止网络攻击扩散的软件，能使检测到计算机病毒后通信切断的应对时间减少至原来的1/30。5月13日，日本《每日新闻》报道，日本正在为其武装部队设计程序，以保护国内核设施不受来自网络空间的打击。5月26日，日本防卫省表示正在研发在服务器遭受攻击而瘫痪的情况下，保证自卫队的指

挥系统仍能维持运行的技术。

(5) 健全网络安全法律

4月11日，日本共同社报道，日本自民党拟定加强网络攻击对策的基本法案框架，规定电力、金融、信息通信等重要基础设施单位有责任和义务维护网络安全。法案还涉及加强“内阁官房信息安全中心”(NISC)的权限及完善体制建设等内容。7月25日，日本政府拟通过一项新法案加强网络信息安全，要求日本各公司不得因负面影响而隐瞒黑客袭击事件，所有黑客事件均需上报。10月14日，日本通过《特定秘密保护法》，公布特定秘密的界定和解除标准，并定于2014年12月10日正式生效。11月6日，日本国会众议院通过了《网络安全基本法》，旨在加强日本政府与民间在网络安全领域的协调和运作，更好地应对网络攻击。

2.2 国外网络信息安全大事件

2.2.1 信息泄密事件

2014年，信息泄露事件频发，个人隐私逐渐成为泄露事件的重灾区。

(1) 斯诺登再曝美国监控内幕

1月，斯诺登曝光美国通过互联网监听从事工业间谍活动。斯诺登称，美国的工业间谍活动所针对的不仅包括国家安全问题，还包括任何可能对美国有价值的工程和技术资料。此后，斯诺登相继又爆出了使用云服务、搜索引擎和社交媒体的有关风险，暗示谷歌和脸谱都与政府勾结进行监听和提供“危险”服务。2月，斯诺登再次曝光美、英、澳、加、新五国间谍机构组成了“五只眼”情报联盟，其间谍行为中包括一个发布虚假信息、操纵网络言论、影响网络安全的“网络魔术师”任务。

(2) 原子能研发机构电脑中毒向韩网站发送信息

1月7日，日本《读卖新闻》报道，日本原子能研发机构6日发布消息称，位于福井县敦贺市“文殊”核电站高速增殖炉作业中心的一台笔记本电脑被发现感染了计算机病毒，相关报告书及邮件内容有可能已经外泄。但泄露的信息均与核研究无关。日本原子能研发机构称，1月2日收到信息安全公司“有不明通信记录”的报告后立即展开了调查。随后发现，在一个半小时里，这台计算机33次向同一个韩国的站点发送信息。

(3) 1600万网民邮箱等信息被盗

1月21日，德国信息技术安全局称，德国有约1600万网络用户的邮箱信息等被盗，作案者为身份不明的黑客。被盗信息主要涉及电子邮件的用户名及密码。许多网络用户不仅在登录邮箱时使用有关信息，在登录社交网站、网购时同样会使用这些信息。

(4) 500万谷歌账户信息被泄露

9月，大约有500万谷歌的账户和密码的数据库被泄露给一家俄罗斯互联网网络安全论坛。这些用户大多使用了Gmail邮件服务和美国互联网巨头的其他产品。据俄罗斯一个受欢迎的IT新闻网站CNews报道，论坛用户tvskit声称60%的密码是有效的，一些用

户也确认在数据库里发现了他们的数据。

(5) 韩国接连发生信息泄露事件

1月，韩国发生大规模个人隐私泄漏事件，韩国国民卡、乐天卡和农协卡等2000万笔信用卡交易用户的个人资料被泄露，网上甚至出现疑似韩国总统朴槿惠、前总统李明博和联合国秘书长潘基文等人士的个人信息。首尔金融监管部门1月19日确认，泄露信息达1亿条以上，受影响的用户数至少达到2000万。

2.2.2 网络攻击事件

2014年，网络攻击事件频发，攻击形式多种多样，攻击的复杂性与频度持续升高，攻击行为已经不单单是黑客单打独斗，而是出现更多错综复杂、有组织性甚至是由敌对国家发起的网络袭击。

(1) 数据库遭黑客攻击，eBay紧急督促1.45亿用户改密码

5月22日，据外媒报道，eBay表示从三个月前开始的黑客攻击已经波及了用户数据库，因此公司督促1.45亿用户修改密码。eBay发言人称大量的账号处于危险中，但是拒绝透露具体的数量。公司经大量测试后发现，没有证据显示攻击造成了未经授权的活动，也没有证据显示财务或信用卡信息遭窃。9月，eBay受到跨站脚本攻击，导致向其部分用户发送恶意网站以窃取用户凭证。有报道称，eBay公司应对安全问题异常缓慢，从首个用户向eBay反映问题开始，网站持续被攻破时间长达12小时。

(2) 比特币交易站受攻击破产

2月，全球最大的比特币交易平台Mt.Gox由于交易系统出现漏洞，75万个比特币以及Mt.Gox自身账号中约10万个比特币被窃，损失估计达到4.67亿美元，被迫宣布破产。这一事件凸显了互联网金融在网络安全威胁面前的脆弱性。

(3) 法国最大电信运营商数据库遭黑客攻击入侵

2月3日，法国最大的电信运营商(Orange)电信公司对外公布，公司的数据库从1月16日开始遭黑客攻击，近80万的客户资料被盗，主要是客户的姓名、邮寄地址，但他们的用户密码并没有被盗。该公司已经阻止了这次黑客攻击并提出控告。法国内情报总局地就此展开了调查。

(4) 美朝两网络对抗趋于公开化

11月22日，美国索尼影视娱乐公司受到自称“和平卫士”的朝鲜黑客组织的攻击，导致公司系统被迫关闭。此次攻击造成包括索尼员工信息、公司计划、产品情况、索尼高层往来邮件、名人电子邮件在内的内部敏感详细信息泄露。12月23日，朝鲜互联网开始出现不稳定状态，使用朝鲜官方域名(.kp)的网站全面陷入瘫痪9小时，26日朝鲜官方通讯社朝鲜中央通讯社网站无法访问持续7小时。事件发生后，美朝两国均否认参与了网络攻击，并均宣称对方政府参与了攻击行为。

2.2.3 重大漏洞事件

2014年是多个网络空间严重漏洞集中爆发的一年，“心脏出血”(Heartbleed)、

“BadUSB”、“破壳”（Shellshock）、“贵宾犬”等重大漏洞先后被曝光，对网站、操作系统、硬件设备的影响范围之广，闻所未闻。

（1）“心脏出血”严重漏洞事件爆发

4月7日，开源软件包 OpenSSL 发现了心脏出血的程序漏洞，在整个 IT 行业及更广的周边行业引起了普遍的恐慌。通过这一漏洞，黑客可以读取到包括用户名、密码和信用卡号等隐私信息在内的敏感数据，并已经波及了大量互联网公司，受影响的服务器数量多达几十万，其中已被确认受影响的网站包括 Imgur、OKCupid、Eventbrite 以及 FBI 网站等。据外媒报道，当前开源软件的使用范围包括国际空间站、股票交易所等重要机构和设施，开源软件的使用比例超过了 95%。

（2）BadUSB 漏洞

8月，在美国黑帽大会上，Jakob Lell 和 Karsten Nohl 公布了 BadUSB 漏洞。攻击者利用该漏洞将恶意代码存放在 USB 设备控制器的固件存储区，而不是存放在其他可以通过 USB 接口进行读取的存储区域。这样，杀毒软件或者普通的格式化操作无法清除该代码，从而使 USB 设备在接入 PC 等设备时，可以欺骗 PC 的操作系统，达到某些目的。

（3）破壳漏洞

9月25日，US-CERT 公布了一个严重的 Bash 安全漏洞（CVE-2014-6271）。由于 Bash 是 Linux 用户广泛使用的一款控制命令提示符工具，从而导致该漏洞影响范围甚广。安全专家表示，由于并非所有运行 Bash 的电脑都存在漏洞，所以受影响的系统数量或许不及“心脏出血”。不过，破壳本身的破坏力却更大，因为黑客可以借此完全控制被感染的机器，不仅能破坏数据，甚至会关闭网络，或对网站发起攻击。

3 国内网络信息安全发展现状

3.1 国内网络信息安全动态

3.1.1 国家网络信息安全动态

2014 年，中央网信办的成立结束了我国网络信息安全部门九龙治水的局面，网络安全工作全面推动。网信办加速了网络信息安全方面立法的出台，推出网络安全审查制度；联合多个部委展开了“打黄打非·净网 2014”、伪基站违法犯罪活动等一系列专项行动；举办了首届世界互联网大会、首届国家网络安全宣传周等系列宣传和普及网民安全意识的活动。

(1) 我国新增七个国家级互联网骨干直联点

1月，工业和信息化部网站发布《关于设立新增国家级互联网骨干直联点的指导意见》，提出除现有北京、上海、广州三个骨干直联点外，在成都、武汉、西安、沈阳、南京、重庆、郑州增设七个新的国家级互联网骨干直联点。国家级互联网骨干直联点（以下简称骨干直联点）作为国家重要通信枢纽，主要用于汇聚和疏通区域乃至全国网间通信流量，是我国互联网网间互联架构的顶层关键环节。建设好、管理好、利用好骨干直联点，关系到我国互联网网络安全、网络效率、产业布局及发展。

(2) 加快建设社会信用体系

1月15日，国务院总理李克强主持召开国务院常务会议，部署加快建设社会信用体系、构筑诚实守信的经济社会环境。会议通过《社会信用体系建设规划纲要（2014—2020年）》，并提出这几点要求。一是全面推进包括政务诚信、商务诚信、社会诚信等在内的社会信用体系建设。政府要以身作则，带头推进政务公开，依法公开在行政管理中掌握的信用信息，提高决策透明度，以政务诚信示范引领全社会诚信建设。二是加强基础建设。制定全国统一的信用信息采集和分类管理标准，推动地方、行业信用信息系统建设及互联互通，逐步消除“信息孤岛”，构建信息共享机制，在保护涉及公共安全、商业秘密、个人隐私等信用信息的基础上，依法使各类社会主体的信用状况透明、可核查，让失信行为无处藏身。三是用好社会力量。企业要把诚信经营作为安身立命之本，切实做到重合同、守信用。发挥行业组织自律和市场机制作用，培育和规范信用服务市场，形成全社会共同参与、推进信用体系建设的合力。四是加快推动立法。把健全相关法律法规和标准体系作为重要基础性工作，列入立法规划尽快推进实施，使信用体系建设有法可依。

(3) 习近平任中央国家安全委员会主席

1月24日，中共中央政治局召开会议，研究决定中央国家安全委员会设置；听取关于一年来贯彻执行中央八项规定情况的汇报，研究部署下一步改进作风工作。中共中央总书记习近平主持会议。会议决定，中央国家安全委员会由习近平任主席，李克强、张德江任副主席，下设常务委员和委员若干名。中央国家安全委员会作为中共中央关于国家安全工作的决策和议事协调机构，向中央政治局、中央政治局常务委员会负责，统筹协调涉及国家安全的重大事项和重要工作。

(4) 《中国互联网站发展状况及其安全报告（2014年）》发布

3月21日，中国互联网协会、国家互联网应急中心发布的《中国互联网站发展状况及其安全报告（2014年）》显示：近年来，我国网站受到的威胁主要包括网页篡改、网站后门、软件漏洞、类型攻击、网页仿冒等类型。在篡改问题上，2013年被篡改的中国网站数量为24 034个，较2012年的16 388个大幅增长46.7%；政府网站是被篡改的主要目标。2013年被篡改的政府网站数量为2 430个，较2012年的1 802个大幅增长34.9%。在植入后门问题上，2013年76 160个中国网站被植入网站后门，其中政府网站有2 425个。后门攻击源主要来自境外IP，其中来自美国20.2%，印尼11.4%，韩国

6.5%，其他32.7%。在网络钓鱼问题上，2013年发现仿冒中国网站的仿冒页面URL地址为30199个，其中美国境内的IP地址承载了12573个针对中国网站的钓鱼页面，占比近42%。被仿冒居前列的包括媒体、银行、互联网社交等网站。

3.1.2 信息企业安全防范动态

2014年信息企业认识到面临的安全问题日益严峻，开始不断加强自身安全问题的防范能力。

(1) 腾讯微信封停3万个假货账号

6月9日，腾讯雷霆行动首次披露了其网络诈骗举报平台上线以来的进展情况：截至2014年6月，微信已累计封停3万个假货公众账号，每日封停1000万条违规和欺诈广告；另外，从3月至6月，QQ及QQ空间针对色情和欺诈行为进行了恶意关系链拆解，拆解量累计超7000万条。

(2) 中文域名“.公司”“.网络”7月开放注册

2014年7月17日，由中国互联网络信息中心(CNNIC)主办的“互联网·中文·新时代”中文顶级域名腾飞计划全球启动仪式在北京举行。启动仪式上，CNNIC负责人李晓东表示，中文顶级域名腾飞计划对推动我国域名行业发展具有重要意义。这一计划将普及中文域名各项应用，助推以“.中国”、“.公司”、“.网络”为代表的中文顶级域名在我国的健康、快速发展，提升中文在世界互联网社群的影响力。

3.1.3 移动互联网安全动态

2014年我国手机网民规模达5.57亿，较2013年增加5672万人。网民中使用手机上网的人群占比由2013年的81.0%提升至85.8%。随着智能手机的普及和手机网民的扩大，移动互联网安全成为传播病毒、木马的重要途径，我国对即时通信工具等也进行了重点整治。

(1) 我国手机用户去年接收3000亿余条垃圾短信

4月11日，北京地区网站联合辟谣平台与搜狗共同发布《2013年度垃圾短信报告》。报告显示，2013年中国手机用户收到的垃圾短信总量超过3000亿条，其中60%的垃圾短信中包含电话号码，20%的垃圾短信包含网址链接，还有一些故意使用错别字或者包含大量标点符号。

(2) 国家三部门启动即时通信工具专项治理行动

5月27日，国家互联网信息办公室联合工业和信息化部、公安部召开专门工作会议，部署在全国范围内开展为期一个月的微信等移动即时通信工具专项治理行动，集中整治移动即时通信公众信息发布服务中的违法违规行为。此次专项行动，重点整治移动即时通信工具的公众平台等公众信息发布服务环节，特别是具有传播和社会动员功能的公众账号。严厉打击移动即时通信公众信息领域传播谣言、暴力、恐怖、欺诈、色情信息等违法违规行为。严厉打击境内外敌对势力的渗透破坏活动。对不认真履行管理责任的企

业依法追究责任。

(3) XX 神器手机病毒爆发，伪装熟人短信传播

8月2日，12321举报中心获悉，国内某公司根据用户举报截获最新安卓手机病毒“XX 神器”。此病毒恶意调取用户短信和联系人权限，发送短信内容：“XX（你的联系人名称），看这个 + 网站地址链接”，单击该链接后会感染病毒，并通过手机通讯录传播病毒信息，微信朋友圈也已开始传播。因为利用熟人网络传播，其扩散趋势和危害程度非常严重。据消息称，深圳网警已于8月2日18时抓获制作传播该病毒的犯罪嫌疑人李某。

(4) 《即时通信工具公众信息服务发展管理暂行规定》

8月7日，国家互联网信息办公室正式发布《即时通信工具公众信息服务发展管理暂行规定》，以十条规定规范以微信为代表的即时通信工具公众信息服务，简称“微信十条”。规定指出，即时通信工具服务提供者应当按照“后台实名、前台自愿”的原则，要求即时通信工具服务使用者通过真实身份信息认证后注册账号。

3.1.4 互联网病毒木马动态

2014年互联网上病毒和木马泛滥、个人信息泄露事件频发，信息安全问题比以往任何一年都更为突出。

(1) 敲诈者病毒

敲诈者病毒会加密硬盘、手机存储卡中的重要文件，在电脑或手机上弹出勒索钱财的提示。一部分敲诈者变种加密的文件可以实现技术解密，另有部分只有病毒作者可以解开，被勒索的网民将面临数据无法访问的结果。

(2) 山寨网银大量窃取网银信息

5月，山寨网银伪装成正常网银客户端的图标、界面，在手机软件中内嵌钓鱼网站，欺骗网民提交银行卡号、身份证号、银行卡有效期等关键信息，同时，部分手机病毒可拦截用户短信，中毒用户将面临网银资金被盗的风险。

(3) 央视报道“危险的 Wi-Fi”

6月，央视《每周质量报告》曝光了黑客通过公共场所免费 Wi-Fi，诱导用户链接从而获取手机中银行卡、支付宝等账户信息，盗取资金的消息，引发了网民对于免费 Wi-Fi 安全性的担忧。

(4) XSS 漏洞

国内知名互联网漏洞平台乌云网表示，招商银行网银存在定向 XSS 漏洞，通杀网页、PC 端及手机 APP，该漏洞可定向窃取信息钓鱼种马。互联网业内人士对此解释称，此漏洞即招行网银某处存储型 XSS 漏洞，黑客可以通过此漏洞，对招行客户进行“钓鱼”、偷密码，并且可以看到账号余额。

(5) SQL 注入漏洞

据乌云网披露，新浪支付系统出现了 SQL 注入漏洞，新浪支付部门已确认该漏洞。

SQL 注入漏洞是指通过控制传递给程序数据库操作语句的关键变量来恶意控制程序数据库，从而获取有用信息或者制造恶意破坏，甚至控制用户计算机系统的漏洞。

3.2 国内网络安全大事件

2014 年，我国网络空间安全事件层出不穷，我们对 2014 年发生的网络空间安全十大重大事件进行了回顾和总结，具体如下：

(1) 我国出现大面积 DNS 解析故障

1 月 21 日，国内通用顶级域的根服务器出现异常，导致我国众多知名网站出现大面积 DNS 解析故障，这一次事故影响到了国内绝大多数 DNS 服务器，近三分之二的 DNS 服务器瘫痪，时间持续数小时之久。事故发生期间，超过 85% 的用户遭遇了 DNS 故障，出现网速变慢和打不开网站的情况，部分地区用户甚至出现断网现象。

(2) 携程漏洞事件

3 月 22 日，有安全研究人员在第三方漏洞收集平台上报了一个题目为“携程安全支付日志可遍历下载，导致大量用户银行卡信息泄露（包含持卡人姓名、身份证号、银行卡号、卡 CVV 码、6 位卡 Bin）”的漏洞。上报材料指出携程安全支付日志可遍历下载，导致大量用户银行卡信息泄露，并称已将细节通知厂商并且等待厂商处理中。

(3) XP 系统停止服务

4 月 8 日，微软公司对 Windows XP 系统停止更新维护的服务。但 XP 仍然是当今世界被广泛使用的操作系统之一。特别是在中国，仍有 63.7% 的用户，也就是大约 3 亿左右的用户还在使用 XP 系统。因此“后 XP 时代”的信息安全一直备受关注。

(4) 中国快递 1 400 万信息泄露

4 月，国内某黑客对国内两个大型物流公司的内部系统发起网络攻击，非法获取快递用户个人信息 1 400 多万条，并出售给不法分子。而有趣的是，该黑客贩卖这些信息仅获利 1 000 元。根据媒体报道，该黑客仅是一名 22 岁的大学生，正在某大学计算机专业读大学二年级。

(5) 大力推动国产化软件发展

5 月 16 日，我国政府采购网公布的《中央国家机关政府采购中心重要通知》称，所有计算机类产品不允许安装 Windows 8 操作系统。7 月，公安部科技信息化局下发通知，称赛门铁克的“数据防泄漏”产品存在窃密后门和高危漏洞，要求各级公安机关今后禁止采购。9 月，银监会正式发布的《应用安全可控信息技术指导意见》中明确指出，从 2015 年起，各银行业金融机构对安全可控信息技术的应用以不低于 15% 的比例逐年增加，直至 2019 年掌握银行业信息化的核心知识和关键技术，安全可控信息技术在银行业达到不低于 75% 的总体占比。这一系列举措意味着我国政府和企业开始正视网络信息安全长期依赖国外技术的现象，国产信息安全软件及企业将迎来新的发展机遇。

(6) 12306 数据泄露事件确认为撞库攻击

12 月，漏洞报告平台乌云网发布一则报告称，大量 12306 用户数据遭泄露，已在互

联网上疯传。泄露的数据包括用户账号、明文密码、身份证号、邮箱等大量用户资料。12306 网站用户数据遭泄露事件引起广泛关注，根据记者从多方得到的综合消息来看，这一事件基本被确认是遭受黑客撞库攻击所致。

4 新技术新应用的发展

4.1 智慧城市的发展

智慧城市是指以物联网、云计算、宽带网络等信息通信技术为支撑，通过信息感知、信息传递及信息利用，实现城市信息基础设施和系统间的信息共享和业务协同，提高市民生活水平和质量，提升城市运行管理效率和公共服务水平，增强经济发展质量和产业竞争能力，实现科学发展与可持续发展的信息化城市。我们认为智慧城市安全存在三方面问题：

(1) 信息基础设施安全的脆弱性增大

在智慧城市大规模的信息基础设施体系中，物联网、超级计算、云计算等关键性技术的商业化应用尚需进行安全和技术论证，网络信息安全前景不明确。如物联网所涉及的关键技术需要统一技术标准以实现兼容，确保所保存和传输信息的机密性、完整性。智慧城市建设体系下复杂的网络接入环境、多样化的接入方式、数量庞大的智能接入终端，需要解决接入时的状态和身份认证，进一步增加了信息安全保障的难度。

(2) 信息安全威胁向城市实体性基础设施领域延伸

智慧城市建设中网络信息安全威胁的非对称性大幅提高。信息基础设施具有全球互联特性，智慧城市中横向关联的信息系统消除了城市各子系统之间的边界，外部主体可以在远距离运用最小的资源实施攻击行为，其身份、位置以及进入路径都难以确认。此外，针对信息基础设施的攻击可以同时迅速传导到金融、能源、交通等部门，侵害实体性基础设施的信息安全，同时导致多个系统瘫痪，损害城市的正常运行，共时性信息安全威胁更为严重。

(3) 城市信息安全责任分担与协同机制复杂

在智慧城市建设中，需要建立一种新的安全责任分担机制，确立政府、私人部门、研究机构等利益相关方的责任并通过法律形式予以明确。政府需要重新界定与私人部门的关系，让相关的私人部门参与到信息安全决策中，因为私人部门不仅是主要的信息使用者，还拥有应对信息安全威胁的技术条件和解决方案。政府作为主要的信息安全管理者，其内部各监管机构之间也需要明确责任。以往城市对基础设施实施监管的路径通常是在建设一种新的基础设施之后，成立一个相应的监管机构，因此城市中存在相互独立的监管电信的机构、监管电网的机构以及监管交通等方面的机构。而智慧城市将实体性基础设施与信息基础设施进行全面互联，需要明确各监管机构在信息安全管理中的职责

及其重要性，设计横向协同机制，以确定各类监管机构的决策层级、可采取的措施及行动边界、协同任务，相关机制设计尤为复杂。

4.2 互联网金融的发展

8月18日，央行发布的第二季度支付体系运行总体情况，全国银行机构处理的电子支付业务（包括网上支付、电话支付和移动支付）为76.96亿笔，金额327.11万亿元，同比分别增长23.24%和3.31%。其中移动支付业务9.47亿笔，金额4.92万亿元，同比分别增长1.55倍和1.37倍，且仍继续保持高速增长。支付机构处理网络支付业务（包括互联网支付、移动电话支付、固定电话支付和数字电视支付）85.32亿笔，金额5.32万亿元，同比分别增长89.43%和129.10%。

互联网金融信息安全部面临的风险主要包括以下几类。

（1）客户端安全认证风险

第一，客户端病毒或黑客入侵。客户端的工作机制是通过用户名和密码相结合进行认证，这种情况下倘若用户被病毒或黑客攻击，安全认证关卡被攻破，那么用户的操作就呈现在黑客的监视下，个人数据会按照黑客指令被发送至病毒或黑客攻击的指定的数据平台或者服务器后端，严重威胁个人账户和密码安全。如：3月16日起，国内最大、最具影响力的P2P网络借贷行业门户网站——网贷之家官网持续多日受到黑客的严重恶意攻击，网站时常无法正常访问，短短几小时内就遭受到6亿次的连续攻击。

第二，假冒网站或诈骗网站。网络域名注册简易，准入门槛较低，虚假恶意网站猖獗。近些年来，有些商业银行、证券公司、投资银行等金融机构均出现过假冒互联网金融的业务，由于这些互联网金融业务的服务器网站域名及网页与真的互联网金融业务相似度极高，一般客户极难辨认。如2012年2~9月，某检测系统检测到的钓鱼网站中有近2/3的钓鱼网站攻击目标对准银行，骗取网民的网银账户信息，对互联网金融信息安全部造成了重大伤害。

第三，垃圾诈骗短信或电话。最常见的是通过拨打电话或发短信等方式，大范围传播各种诈骗信息或垃圾广告，窃取用户钱财进行非法牟利。通常情况下，该类型事件的发生皆因手机信息在网上泄露而生，且与信息安全高度相关。

（2）技术安全风险

第一，技术应用风险。技术应用风险是由于最初设计构思的片面性、局限性等导致的互联网金融技术系统存在明显缺陷而产生的风险。据国家互联网应急中心2014年2月的数据显示，我国境内网站数量被篡改数量高达12428个，220万余个终端感染病毒，另外泄露信息系统安全的漏洞高达699个。

第二，技术能力风险。技术能力风险是指因为互联网金融平台存在固有的技术缺陷，由于这些固有的技术缺陷在某些特定情况下导致的无可避免的风险。2011年EnfoDesk易观智库《中国第三方网络支付安全调研报告》数据显示，用户在互联网金融使用过程中由于木马、钓鱼网站和账户、密码被盗等因素带来资金损失所占的比例是最高的，分别

为 24% 和 33.9%。

(3) 数据安全风险

在大数据、云计算、物联网等环境下，互联网金融的数据安全将面临更大的风险挑战。互联网金融机构掌握着客户的需求偏好、投资定位、贷存款、信用情况等信息，一旦这些个人信息遭到不法分子窃取、泄露、非法篡改等，将会对客户隐私、投资者权益以及信用情况等个人隐私构成严重威胁。如 2011 年，网友爆料几家银行均泄漏几千万用户数据，包括银行卡号、密码、个人基本信息等，并配图证实，此消息一经发布，在网络迅速蔓延，造成公众极大恐慌。银行、互联网金融机构以及用户三者之间通过互联网通道进行数据传输是互联网金融业务的工作方式，这三者在数据传输过程中必须要对数据进行加密。但目前加密算法和传输系统安全性还不是特别完善，存在网络传输系统和环境被击破，或者加密算法被黑客攻破的风险，这种情况一旦发生，客户的各种交易操作就暴露于众目睽睽之下，信息安全将受到极大威胁。

4.3 社交网络的发展

社交网络是继门户网站、搜索引擎之后，互联网发展的第三次浪潮。互联网时代的社交网络则给人类社会生活带来革命性的变化。社交网络面临的威胁主要分为三类：用户隐私相关的安全威胁、传统的网络与信息安全威胁、身份相关的安全威胁。

(1) 用户隐私相关的安全威胁

据最新研究表明，攻击者主要通过公开的名单列表、虚假用户、合法用户、钓鱼攻击、恶意软件、面部识别、基于内容的图像检索、图像元数据标记等方式，从社交网络中有效获取大量个人信息。

1) 公开的名单列表。利用爬虫技术来大规模获取用户数据，目前已有专业的提供商来向用户提供付费爬虫服务，下载公开可得的用户信息。

2) 虚假用户。通过合法的邮件地址来申请虚假用户，通过建立大量的虚假用户来大规模发展好友，以此获取用户的个人信息。

3) 合法用户。通过非法手段来窃取合法的用户账号，特别是那些具有相当规模好友链的用户，如“万人迷”用户，从而达到迅速获取大量用户信息的目的。

4) 钓鱼攻击：伪造 URL 链接，将链接指向虚假网站。社交网站经常向用户邮箱发送好友状态更新的 URL 链接，然而普通用户很难区分这些 URL 链接的合法性，故非常容易受到钓鱼攻击的威胁。

5) 恶意软件：许多社交网络向第三方提供了应用程序开发接口，使得平台上的第三方软件具有读取用户重要数据信息的权限。由于用户未对这些软件的访问权限给予限制，许多恶意软件在提供正常功能的同时，也可以收集用户数据信息，达到窃取信息的目的。

(2) 传统的网络与信息安全威胁

传统的网络与信息安全威胁在社交网络环境下仍十分明显，主要表现为垃圾信息、跨站脚本。下面将分别介绍。

1) 社交网站垃圾信息, 即通过社交网站传播的用户未请求接收的信息。在某些领域, 社交网络已取代 E-mail 成为人们主要的通信工具。这意味着, 原来影响 E-mail 的同等规模的垃圾信息现在又将影响社交网络。攻击者利用社交网站快速增长的用户数量传播信息。有统计表明, Facebook 用户间通信的消息中, 有 43% 的消息被认为是垃圾信息。垃圾信息对社交网站造成的危害极大, 如流量拥塞、社交网络信任缺失、色情链接泛滥等。

2) 跨站脚本, 即用户在个人信息空间及留言板中使用的各类脚本代码, 如 flash、html 等。众多的社交网络都为第三方开发者提供免费的应用程序开放接口, 因此攻击者可利用跨站脚本进行攻击。而且, 由于社交网络中用户间的消息传递十分迅速, 使得病毒极易在用户间扩散, 如 SAMY 病毒曾在不到 20 小时内就感染了 100 万 MySpace 用户。Koobface 蠕虫病毒可在 Facebook 内自动发出好友邀请, 窃取用户密码等。

(3) 身份相关的安全威胁

身份相关的安全威胁是指攻击者利用社会网络中的用户信息从事各类违法活动, 包括使用社交网站的鱼叉钓鱼攻击、面向社交网站的钓鱼攻击、网络渗透和用户空间抢注与名誉诽谤等。

1) 社交网站的鱼叉钓鱼攻击, 即一种高度目标化的钓鱼攻击。社交网站的鱼叉钓鱼攻击不同于普通的利用邮件的攻击, 而是利用在社交网站上获得的信息, 来进行目标性很强的钓鱼攻击。相比于普通钓鱼攻击 15% 的成功率, 利用社交网站上的公开信息实施的钓鱼攻击, 成功率可达到 72%。

2) 面向社交网站的钓鱼攻击, 即指钓鱼攻击就在社交网站上进行。面向社交网站的钓鱼攻击并不将从社交网站上获取的信息用于他处, 而是以社交网站为媒介来进行的攻击, 如在发送给好友的链接中添加恶意代码。此类攻击主要用于身份盗用、名誉毁谤等。

3) 网络渗透, 即攻击者渗透至社交网络用户的好友圈。因为通常用户的某些信息只是允许好友才能访问的, 用户的好友圈或特定用户群是防范个人信息泄露的第一道防线。然而这到防线很容易被攻破。多数社交网站的目的是发展尽可能多的用户来提高它们的流量与广告收益, 所以大多在加入好友申请时都没有验证码功能。这为通过脚本添加好友, 进行网络渗透提供了便利。

4) 用户空间抢注与名誉诽谤, 即在社交网站中假冒著名的公众人物或公司来对他们进行诽谤。通过身份盗用实施的用户空间抢注与名誉诽谤, 一旦在社交网站盗用了别人的身份, 则用户相应的身份及好友信息全部泄露无疑, 给攻击者实施钓鱼攻击或名誉诽谤创造了条件。

5 对策和建议

5.1 智慧城市安全对策

我国的智慧城市发展还有很长的路要走, 未来我国智慧城市的发展对策和建议主要

体现在以下几个方面：

(1) 加强智慧城市顶层设计

从目前我国智慧城市的发展可以看出，各个地方政府对智慧城市的发展相当混乱，只有加强对智慧城市的顶层设计，才能对智慧城市的建设做出合理的规划。十八大以来，党和国家提出了促进新型城镇化、全面建设小康社会的方针战略，而智慧城市的建设也要形成相应的方针和这些政策同步进行，只有把智慧城市的理念融入到大的政策方针中，才能使智慧城市的建立更加有序合理，这个理念不仅可以在城市推广，甚至可以推广到城镇和农村，从而卷起全国性的智慧城市建设浪潮。

(2) 建立综合协调机制

智慧城市的建设涉及的领域非常多，其中包括信息技术、医疗卫生、城市交通、教育和社区管理等，这些领域相互独立，因此要发展智慧城市，必须要建立一个统筹机制对这些领域进行协调。但在我国现有的发展水平，尚且没有这样的一个统筹机制进行相应协调。最初智慧城市的提出和建设是信息化主管部门一手管理，但随着不断发展，城市规划建设部门也对智慧城市的发展制定了相应的目标战略，最近住房和城乡建设部也参与其中，这三个部门的发展方向不尽相同，以信息化主管部门主导的智慧城市能够很好地将新一代信息技术运用于智慧城市的建设当中，但往往不能把规划落实到具体城市建设项目中，而城市规划部门虽然能将制定的规划落实到具体建设项目中，但很少能将新一代信息技术进行融合，这就造成了智慧城市建设不能全面发展。

(3) 建立信息网络安全机制

我国在网络信息安全建设方面比较落后，很多信息网络安全问题都由外国，尤其是美国的公司进行技术支持，这使得我国的网络安全风险大大增加，尤其是随着云计算、物联网等新一代信息技术的发展，这种安全问题表现得更加突出，为了保障我国信息化的发展，必须建立一个信息网络安全机制。具体来说可以从实施和管理两方面入手，一方面要保障智慧城市所采用的网络设备和基础数据库的安全性，另一方面要建立各种保障机制，包括国家网络安全评估机制、入侵监测和防范体系、技术产品资质认定和采购备案机制、安全监测和应急反应机制，通过对信息网络安全的评估、对设备提供企业资质的审查和对外网的监测，从而保障信息网络的安全性，为智慧城市的建设和发展提供有力保障。

(4) 加强信息安全建设

智慧城市虽然采用了诸多新型信息技术，改变了信息服务方式，但并没有颠覆传统的信息安全模式。我们可以通过分析智慧城市总体框架中的应用层、数据层、通信层、感知层、安全体系等多个层面的实际安全需求，结合产品本身的技术特点和智慧城市中所采用新型技术的具体情况，以“智慧监测、智慧防护、智慧审计、智慧应用”四大安全途径为基础，形成一套完善的针对智慧城市安全事件的综合管理平台解决方案，以保障智慧城市信息系统的安全、稳定、高效、发展。

5.2 互联网金融安全对策

网络金融安全涉及多方面的内容，针对目前我国金融安全方面存在的问题，可以采

取相关的措施加以解决，保障我国网络金融的安全。具体来讲，可以从以下几个方面着手。

(1) 明确金融主体，建立有效的监管体系

互联网金融虽然是互联网与金融业深度合作发展的新兴行业，但其本质仍是金融行业，从事互联网金融的机构应属于金融机构，纳入金融体系管理。我国传统金融业经过多年累积和发展已经建立起了一套较为规范、完整的监管体系，而新兴的互联网金融应借鉴传统金融业的监管体系并结合自身的特殊性制定相应的监管规则，使互联网金融能够在一定监管下健康有序地发展。由于互联网金融呈现的是混业经营态势，传统的分业监管模式已无法满足其监管要求，因此从改善市场环境和促进行业健康发展的角度，建议推进以人民银行牵头的综合监管，即根据互联网金融所涉及的领域建立以监管主体为主、其他相关部门相互协调配合、行业自律为补充的新的合作监管体系。

(2) 互联网金融与传统金融合作共赢

互联网金融与传统金融业在相互博弈中应取长补短，形成全方位覆盖、多样化服务、多层次体系的金融格局。从目前互联网金融的发展形势来看，一方面，在短期内传统金融业无法排挤互联网金融，单从技术层面上来说，互联网金融利用其特有的现代信息技术，不仅弱化了传统金融中介的作用，加速金融脱媒，更重要的是给参与者带来了快捷、方便和高效的服务。而传统金融业要想获得长远发展，必须依赖互联网等高科技拓展金融业务，丰富金融产品，并利用互联网的先进技术来实现自我优化升级和弥补自身缺陷。另一方面，互联网金融在短期内也难以替代传统金融业，互联网金融的应用领域先天受技术限制，发展规模后天受国家政策影响，使其无法发挥潜在的优势。更重要的是，传统商业银行在支付体系中不可或缺的位置及在存贷款领域的强势地位对于互联网金融企业来说都是无法企及的。

(3) 建立互联网金融安全制度，防范系统性技术风险

构建互联网金融安全制度不仅要求互联网金融企业改进其运行环境，对顾客的交易资料加强管理，更重要的是要开发自主产权设备。首先，改进互联网金融运行环境应从硬件和网络运行两方面入手：在硬件上要增强对硬件安全措施的投入，提高计算机系统的防病毒能力，确保硬件环境顺利运行；在网络运行上，为了限制非法用户的登录应采用分级授权和身份认证来登录。其次，在加强客户资料管理方面：一方面要积极整合各种互联网资源，建立以客户为中心的互联网金融数据库，通过分析数据库中的业务信息实现对互联网金融业务流程的全程监控；另一方面要加大开发加密技术、密钥管理技术及数字签名技术来降低技术选择风险。第三，要大力支持自主知识产权的信息技术研发，力求在数据加密技术、防火墙技术等网络安全技术方面有重大突破，早日实现我国互联网金融技术上的独立，这样才能使我国互联网金融稳健、持续地发展。

(4) 创新互联网金融盈利模式

互联网金融在我国已发展多年，但无论是第三方支付还是网络贷款，面对市场和政策的双重压力，其业务对象只是那些被传统金融忽略的客户，且主要的盈利模式是利用平台向这些客户收取一定的手续费或中介费。然而，当余额宝等互联网金融创新产品横

空出世，银行等传统金融业也开始对互联网金融业务产生兴趣，于是出现了新一批“宝宝”类的金融产品，这样就使得互联网类金融产品同质化更加严重。鉴于此，互联网金融企业要想在竞争中立于不败之地，必须要探索新的业务领域，不断创新金融产品和金融服务，充分利用互联网技术与传统金融业形成差异化的竞争格局。同时，互联网金融企业还要根据用户的需求，准确定位最能满足客户需求的金融产品和服务，最终靠低成本、优服务、高质量来赢得大众的青睐。

5.3 社交网络安全对策

据 360 安全中心发布的《2015 年第一季度中国互联网安全报告》显示，社交工具和电子商务网站是网络诈骗信息传播的最主要途径，占比均为 37.2%。超过了搜索引擎、分类信息网站等网络诈骗信息传播途径。社交网络正在成为黑客挖掘的“金矿”。如何解决社交网络用户信息安全保护中诸多的不足与缺陷，需要业界与研究者不断探究与摸索。

(1) 加强立法保障和行业自律

我国需要借鉴其他国家的经验，尽快进行个人信息安全保护立法，科学地设置个人信息安全保护的相关细则，明确规定个人信息安全的界限、信息主体的权利和义务等内容。比如，中国互联网协会、中国电子商务诚信联盟等行业组织都已积极主动地承担起责任，2002 年 4 月公布施行的《中国互联网行业自律公约》第八条规定：“自觉维护消费者的合法权益，保守用户信息秘密；不利用用户提供的信息从事任何与向用户做出的承诺无关的活动，不利用技术或其他优势侵犯消费者或用户的合法权益。”还应该成立专门的网络用户信息安全保护机构，制定统一的网络认证体系，对社交网络进行监控，对符合规定的网站颁发“营业执照”，并定期地对获得执照的网站进行隐私保护程度等級评定，张贴不同等级的徽章或记号，此评定为所有用户清晰可见，表明网站的可信任程度，以此来促进网站加强自我管理。另外，不管是主管部门还是行业协会，都应该充分利用各种新闻媒体广泛宣传网络用户信息安全保护的重要性，教育用户主动了解各种法律法规等。

(2) 提高社交网络企业信息安全管理水

无论是国家立法还是行业自律，都只是依靠外部力量进行规范，社交网络企业要提高对用户信息安全的保护水平。用户信息安全保护的主体是广大社交网络企业。他们出于开展业务的需要而收集大量用户信息，因此广大社交网络企业要深刻意识到用户信息安全的重要性，把用户信息安全作为企业信息安全的重要组成部分，加大信息安全建设与管理的力度。首先，社交网络企业的网站安全性至关重要，目前一些社交网站大量使用了 Ajax 技术，必须对 XSS（跨站脚本攻击）和 CSRF 攻击以及相关的蠕虫病毒进行关注，在网站建设初期加强对用户输入内容可靠性的限制管理，同时还要做深入细致的检测，最基本的就是对有输入框的页面进行代码测试。其次，应该有良好的信息伦理道德，强化对用户信息安全进行保护的意识，梳理完善用户信息安全管理规章制度和流程，细化用

户隐私保护条款，出台网站安全访问指南，在网站醒目位置提醒用户注意保护个人信息。

(3) 提升社交网络用户信息安全素养

用户信息安全素养的提高对信息安全保护至关重要。首先，用户要主动学习信息安全有关知识，如有可能，参加相关培训班，提升自我信息安全意识；如果用户是工作人员，则要遵守所在单位颁布的社交使用守则；其次，应对社交网站情况有一个全面的考察，主动了解网站收集使用个人信息的目的、用途等，不应认为网站有了隐私保护条款，就能对用户的信息安全提供全方位的保护，有些网站的隐私保护条款的制定只是为了赢得用户的信任或者迫于压力而不得不制定；第三，用户应自行采取技术手段设置权限，禁止 Cookies 等的应用，对自己发布的个人信息要深思熟虑，不要填写过于详细的个人信息，比如收入水平、婚姻状况等，更不得随意泄露个人信息，以防被不怀好意的人利用，进行诈骗或者商业推广；最后，要做好个人计算机的杀毒、防木马等安全工作，定期更新杀毒软件。

6 结束语

2015 年是全面深化改革的关键之年，是全面依法治国的开局之年，是我国网络安全工作励精图治、深化改革、锐意创新的关键之年。我们应深入贯彻党的十八大，十八届三中、四中全会精神，按照习近平总书记提出的建设网络强国，坚决打赢网上斗争的战略目标，以及网络安全和信息化必须统一谋划、统一部署、统一推进、统一实施的明确要求，统筹国内、国际两个大局，加强顶层设计和协调配合，才能共同有效维护国家网络空间安全。

参考文献

- [1] 张颖. 计算机网络的信息体系结构[J]. 电子制作, 2015(3).
- [2] 祖金红. 网络环境下的计算机信息处理与安全技术探讨[J]. 电子制作, 2015(3).
- [3] 王世伟. 论信息安全、网络安全、网络空间安全[J]. 中国图书馆学报, 2015(3).
- [4] 谭显红, 晏茂楠. 网络信息安全的技术分析与防护措施[J]. 电子技术与软件工程, 2015(6).
- [5] 张海旭. 关于网络信息安全[J]. 电子技术与软件工程, 2015(7).
- [6] 李艳. 当前国际互联网治理革新动向探析[J]. 现代国际关系, 2015(4).
- [7] 朝闻. 网络安全追击[J]. 互联网周刊, 2015(4).
- [8] 黄赛. 移动互联网时代手机信息安全策略[J]. 无线互联科技, 2015(4).
- [9] 许俊良. 信息化建设的作用以及云计算对我国的信息化推动[J]. 信息系统工程, 2015(6).
- [10] 韩全惜. 论网络信息安全技术防护体系建设方法[J]. 无线互联科技, 2015(4).
- [11] 于雪莲. 电子政务系统中信息安全技术研究与应用[J]. 信息系统工程, 2015(6).
- [12] 李铁军, 刘斌. 计算机通信与网络发展应用技术浅谈[J]. 信息系统工程, 2015(6).

- [13] 陈飞. 互联网产业投资模式探析[J]. 宏观经济管理, 2015(7).
- [14] 姚相振. 2015 年 RSA 大会网络安全热点议题[J]. 信息技术与标准化, 2015(6).
- [15] 黄阳华, 林智, 李萌. 互联网 + 对我国制造业转型升级的影响[J]. 中国党政干部论坛, 2015(6).
- [16] 贾楠. 我国互联网金融监管创新问题研究[J]. 北京市经济管理干部学院学报, 2015(6).
- [17] 周汉民. 互联网金融法治当先行[J]. 上海市社会主义学院学报, 2015(7).
- [18] 郝叶力. 大国网络战略博弈与中国网络强国战略[J]. 国际关系研究, 2015(7).
- [19] 崔久强, 徐祺. 移动互联网身份认证技术研究[J]. 信息安全与技术, 2015(7).
- [20] 罗拥华, 邱尚明, 姚幼敏. 云计算背景下计算机安全问题及对策[J]. 电子制作, 2015(7).
- [21] 柳劲华. 网络安全策略[J]. 信息安全与技术, 2015(7).
- [22] 张汉卿. 计算机网络安全问题及其防范措施[J]. 科技展望, 2015(7).

作者简介

郝文江 副研究员, 就职于公安部第一研究所, 主要从事信息安全、网络安全、计算机犯罪侦查取证方面研究, 主持开展了国家级项目 1 个、省部级项目 5 个, 参与国家级和省部级项目 4 个, 已在国内外学术刊物上发表论文 50 余篇。



徐丽萍 工程师, 就职于公安部第一研究所, 主要从事信息安全、网络安全等方面研究。



李翠翠 助理工程师, 就职于公安部第一研究所, 主要从事信息安全、网络安全方面研究。



大数据机器学习的研究进展与趋势

CCF 人工智能与模式识别专委会

摘要

本报告简要介绍国内外大数据机器学习的研究进展，重点关注以哈希学习、随机投影为代表的数据消减技术和以分布式学习、随机优化为代表的算法伸缩技术，并展望未来的发展趋势。

关键词：机器学习，大数据机器学习，大数据学习，哈希学习，随机投影，分布式学习，随机优化

Abstract

Big data machine learning has become an indispensable tool for discovering knowledge from data. In this paper, we provide a brief survey of recent developments in big data machine learning, focusing on data reduction techniques such as learning to hash and random projection and scalable techniques such as distributed learning and stochastic optimization. Furthermore, some future research directions in this area are also discussed.

Keywords: big data, big data machine learning, big learning, learning to hash, random projection, distributed learning, stochastic optimization

1 引言

众所周知，大数据时代已经来临^[1]，各行各业积累的数据都呈现出爆炸式增长趋势。例如，欧洲粒子物理研究所（CERN）每秒产生 40TB（1TB 等于 1 024GB）的数据，Facebook 每天处理的数据超过 10TB，淘宝网每天的登录用户大约有 6 000 万，页面浏览量约 20 亿次，新浪微博每天上传的微博数超过 1 亿条。中国移动研究院的一份简报称，2011 年人类创造的数据达到 180 亿 GB，而且每年还在以 60% 的速度增长，预计到 2020 年，全球每年产生的数据将达到 350 万亿 GB。大数据对人类社会的各方面都带来了巨大的影响甚至是变革，例如在科学领域，图灵奖得主 Jim Gray 明确指出，数据密集型科学发现已经成为科学研究中在实验探索、理论分析、计算验证之后的第四范式^[2]，“数据科学”应运而生。

收集、传输、存储大数据的目的，是为了“利用”大数据，若没有机器学习技术来分析大数据，“利用”则无从谈起^[3]。从数据分析的角度看，大数据带来了很多挑战和机遇^[4]。大数据机器学习（简称大数据学习）可以为大数据智能分析提供核心技术支撑，引起了研究者的广泛关注^[5]。例如，顶级国际会议 NIPS 从 2011 年开始多次举办关

于大数据学习的专题研讨会 (<http://biglearn.org/>)，组织者包括多名机器学习领域的国际知名学者，并吸引了企业界和学术界的广泛参与。奥巴马政府提出“大数据计划”后，美国 NSF 在加州大学 Berkeley 分校启动了加强计划，专门研究如何整合美国 NSF 所认为的大数据时代的三大关键技术——机器学习、云计算、众包，而加州大学 Berkeley 分校于 2013 年特别举行了大数据学习的学术研讨会 (<http://simons.berkeley.edu/programs/bigdata2013>)，组织者包括美国科学院、工程院两院院士 Michael Jordan，以及美国工程院院士 Stephen Boyd 等大数据学习领域的知名学者。

大数据具有很多特点，例如被经常提及的“4V”：Volume（巨量）、Velocity（快速）、Variety（多样）和 Value（高值）。大数据机器学习涉及机器学习研究的理论、方法、应用等所有方面。广义来说，现有的机器学习研究都在从不同角度和层面触及到有关大数据的问题，因此都可谓之大数据学习；但由于大数据涉及的机器学习问题非常丰富、复杂，作为一个新起步的研究领域，狭义的大数据学习目前主要关注的是如何更好地对巨量数据进行学习，因此，本报告所讨论的大数据机器学习主要指近几年出现的、为解决数据的巨量性而提出的新型机器学习模型与算法。

针对数据的巨量性，最近的研究围绕着模型构建与模型学习两个方面展开。在模型构建方面，大数据的出现使得复杂模型的有效训练成为可能。因此，目前大数据机器学习的一个重要研究内容是设计有效的复杂模型，其中最具代表性的是深度学习模型，这方面的内容可以参考中国计算机学会（CCF）人工智能与模式识别专业委员会撰写的《CCF 2013—2014 中国计算机科学技术发展报告》的相关报告^[6]，本文不再赘述。本文主要关注模型学习方面的内容，即当数据量大到目前已有的学习算法不能处理的规模时，怎样设计有效的机制来实现对模型的高效学习。

解决该问题一般有两种思路：一种思路是对数据进行消减，减少数量或降低维度，使其变成传统算法能处理的规模；另一种思路是对现有算法进行改造，设计可伸缩（扩展）的学习算法。在数据消减方面的技术主要包括哈希学习、随机投影等技术，这些技术能减小存储和通信开销，提高大数据学习系统的效率。算法伸缩方面的技术主要包括分布式学习和随机优化，这些技术能实现学习算法的快速计算，可扩展性强。

2 国际研究进展

本节将围绕数据消减与算法伸缩两个方面对大数据机器学习的国际研究进展进行介绍。

2.1 数据消减

2.1.1 哈希学习

哈希学习（learning to hash）^[7-15]通过机器学习机制将数据映射成二进制串的形式，

能显著减少数据的存储和通信开销，从而有效提高学习系统的效率。哈希学习的目的是学到数据的二进制哈希码表示，使得哈希码尽可能地保留原空间中的近邻关系，即保相似性。具体来说，每个数据点会被一个紧凑的二进制串编码，在原空间中相似的两个点应当被映射到哈希码空间中相似的两个点。图 1 是哈希学习的示意图，以图像数据为例，原始图像表示某种经过特征抽取后的高维实数向量，通过从数据中学习到的哈希函数 h 变换后，每幅图像被映射到一个 8 位 (bit) 的二进制哈希码，原空间中相似的两幅图像将被映射到相似（即海明距离较小）的两个哈希码，而原空间中不相似的两幅图像将被映射到不相似（即海明距离较大）的两个哈希码。

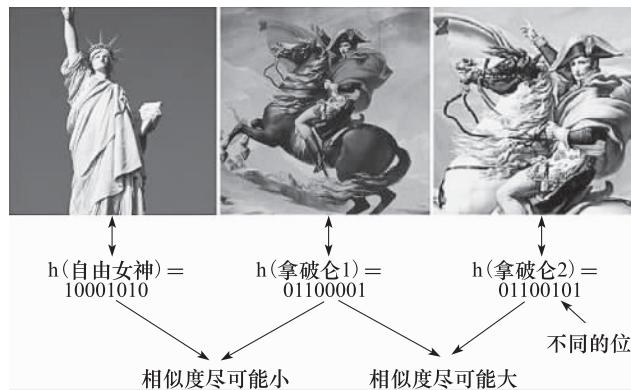


图 1 哈希学习示意图

使用哈希码表示数据后，所需要的存储空间会被大幅减小。举例来说，如果原空间中每个数据样本都被一个 1 024 字节的向量表示，一个包含一亿个样本的数据集要占用 100GB 的存储空间。相反，如果把每个数据样本哈希到一个 128 位的哈希码，一亿个样本的存储空间只需要 1.6GB。单台机器（包括配置很高的单台服务器）处理原始表示时，需要不断地进行外内存交换，开销非常大。但如果用哈希码表示，所有计算都可以在内存中完成，单台普通的个人电脑（PC）也能很快地完成计算。由于很多学习算法，比如 k 近邻（kNN）、支持向量机（SVM）等的本质是利用数据的相似性，哈希学习的保相似性将在显著提高学习速度的同时，尽可能地保证精度。另一方面，因为通过哈希学习得到的哈希码位数（维度）一般会比原空间的维度要低，哈希学习也能降低数据维度，从而减轻维度灾难问题。此外，基于哈希学习得到的二进制哈希码可以构建索引机制，实现常数或者次线性级别的快速近邻检索，为上层学习任务的快速实现提供支撑。因此，哈希学习在大数据学习中占有重要地位。

需特别指出的是，数据库研究领域早已使用二进制哈希码来表示数据^[16-18]，但他们使用的哈希函数是人工设计或者随机生成的；与之不同，哈希学习是希望从数据中自动地学习出哈希函数。从哈希技术的角度来看，前者被称为数据独立方法，后者被称为数据依赖方法。有研究表明^[12-13]，与数据独立方法相比，数据依赖方法（即哈希学习方法）只需用较短的哈希编码位数就能取得理想的精度，从而进一步提高检索和学习效率，降低存储和通信开销。

哈希学习由 Salakhutdinov 和 Hinton^[7-8]于 2007 年推介到机器学习领域，于近几年迅速发展成为机器学习领域和大数据学习领域的一个研究热点^[9-15, 19-22]，并广泛应用于信息检索^[23]、数据挖掘^[24]、模式识别^[25]、多媒体信息处理^[26]、计算机视觉^[27]、推荐系统^[28]，以及社交网络分析^[29]等领域。

由于从原空间中的特征表示直接学习得到二进制的哈希编码是一个 NP 难问题^[9]，现在很多的哈希学习方法^[9, 12-14]都采用两步学习策略：第一步，先对原空间的样本采用度量学习（metric learning）^[30]进行降维，得到一个低维空间的实数向量表示；第二步，对得到的实数向量进行量化（即离散化）得到二进制哈希码。现有的方法对第二步的处理大多很简单，即通过某个阈值函数将实数转换成二进制位。通常使用的量化方法为一个阈值为 0 的符号函数，即如果向量中某个元素大于 0，则该元素被量化为 1，否则如果小于或等于 0，则该元素被量化为 0。例如，假设样本在原空间中的特征表示为一个 5 维实数向量（1.1, 2.3, 1.5, 4, 3.2），经过某种度量学习（通常把降维看成度量学习的一种）处理后得到一个 3 维的实数向量（1.8, -2.3, 0.6），然后经过符号函数量化后，得到的二进制哈希码为（1, 0, 1）。一般来说，度量学习阶段首先得构建学习模型，然后对模型的参数进行优化和学习。下面我们将从学习模型、参数优化和量化策略三方面来介绍哈希学习的最新进展。

根据学习模型（一般指度量学习阶段的模型）是否利用样本的监督信息（例如类别标记等），现有的哈希学习模型可以分为非监督模型^[13-14]、半监督模型^[12]和监督模型^[19, 22, 25]。非监督模型又可以进一步细分为基于图的模型^[14]和不基于图的模型^[13]，监督模型又可以进一步细分为监督信息为类别标记的模型^[19, 25]和监督信息为三元组或者排序信息的模型^[22]。实际上，这每一个细分的类对应于机器学习中一个比较大的子方向，例如基于图的模型。由此可以看出，现有的哈希学习模型虽然总数比较多，但是在各个子方向上还仅仅只是进行了初步的尝试。此外，度量学习是机器学习领域的研究热点之一，而度量学习方面的工作刚好可以用来实现哈希学习的第一步，因此目前很多哈希学习模型（包括非监督、半监督和监督）只是直接利用或者简单改进已有度量学习模型，然后采用上述的符号函数进行量化，得到哈希编码。度量学习得到的结果通常是在模型目标函数的限制下使得信息损失最小，因此得到的总是最优的结果；而在将度量学习应用到哈希学习中时，除了第一步的度量学习可能造成信息损失外，第二步量化过程的信息损失对性能的影响也非常大，有时候甚至超过第一步造成的信息损失。因此，第一步度量学习得到的最优结果并不能保证最终量化后的二进制编码为最优。目前，很多哈希学习方法没有将量化过程中的信息损失考虑到模型构建中去。

现有的参数优化方法大概可以分为两类。第一类是采用与传统度量学习的优化方法类似的策略，对所有位对应的（实数）参数一次性全部优化^[9, 14]。这种策略带来的一个不利后果是没办法弥补量化过程带来的信息损失，有可能导致的结果是随着哈希码长度的增大，精确度反而下降。第二类是避免一次性全部优化所有位对应的（实数）参数，而采用按位（bitwise）优化策略^[12]，让优化过程能够自动地弥补量化过程中损失的信

息。实验结果表明，即使学习模型的目标函数相同，采用按位优化策略能取得比一次性全部优化所有参数的策略更好的性能。但按位优化策略对模型目标函数有一定的要求和限制，比如目标函数可以写成残差的形式。目前，大部分哈希学习方法还是采取一次性全部优化所有参数的策略。

哈希学习跟传统度量学习的一个很本质的区别是需要量化成二进制码。现有的哈希学习方法大多采用很简单的量化策略，即通过某个阈值函数将实数转换成二进制位。最近出现一些专门研究量化策略的工作，并且发现量化策略也会影响哈希学习方法的性能，至少跟第一步的度量学习阶段同等重要。一般来说，度量学习的结果中，各维的方差（或信息量）通常各不相等^[13]。而现有的很多方法采用“度量学习 + 相同位数编码”的策略^[9,14]，导致的结果是随着哈希码长度的增大，精确度反而下降。一种更合理的量化策略是，采用更多的位数编码信息量更大的维。目前，有部分工作在这方面进行了尝试，取得了不错的结果^[31]。

2.1.2 随机投影

随机投影（random projection）通过随机的方式产生投影方向，达到降低数据维度的目的^[32]。随机投影简单快速，适用于大数据场景，并且具备良好的可分析性，利用随机过程等数学工具能够建立坚实的理论基础。

给定一组 d -维的数据 $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbf{R}^d$ ，随机投影首先产生一个随机矩阵 $R \in \mathbf{R}^{d \times m}$ ，其中 $m \ll d$ ；然后利用 R 将数据投影到 m -维空间中：

$$\hat{\mathbf{x}}_i = R^T \mathbf{x}_i, \quad i = 1, \dots, n \quad (1)$$

这样，后续的机器学习任务，如聚类和分类，只需要处理一组 m -维的数据 $\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_n \in \mathbf{R}^m$ ，大大降低了计算量^[33,34]。从上面的流程可以看出，随机投影的计算复杂度主要集中在(1)中的投影计算，其复杂度为 $O(mnd)$ ，生成投影矩阵的计算复杂度 $O(md)$ 并非计算瓶颈。原因在于随机投影是一种与数据独立的降维方法，投影矩阵的生成与数据量 n 没有直接关系。与其相对应的是以主成分分析、流形学习为代表的数据依赖的降维算法。一般而言，数据依赖的降维算法效果更好，但存在计算量过大的问题。比如主成分分析生成投影矩阵的计算复杂度为 $O(nd^2 + d^3)$ ^[35]，对数据量 n 和原维度 d 有较高的依赖。

尽管随机投影非常简单，但理论分析表明随机投影具有非常好的性质，比如随机投影能够保持数据之间的欧氏距离^[36-38]、内积^[38]、体积和到仿射空间的距离^[39]。这些良好的性质是随机投影被广泛应用的理论基础。下面我们简单介绍一下随机投影能够保持欧式距离这一性质。根据^[38]中的定理 2，我们有如下结论。

定理 假设随机矩阵 $R = S/\sqrt{m}$ ，其中 $S \in \mathbf{R}^{d \times m}$ 的元素是从高斯分布 $\mathcal{N}(0, 1)$ 或均匀分布 $\mathcal{U}(-1, 1)$ 中独立采样得到。对于任意两个点 $\mathbf{x}, \mathbf{y} \in \mathbf{R}^d$ ， $0 < \epsilon < 1$

$$(1 - \epsilon) \|\mathbf{x} - \mathbf{y}\|_2^2 \leq \|R^T \mathbf{x} - R^T \mathbf{y}\|_2^2 \leq (1 + \epsilon) \|\mathbf{x} - \mathbf{y}\|_2^2 \quad (2)$$

成立的概率大于 $1 - 2\exp\left(-\frac{m}{4}(\epsilon^2 - \epsilon^3)\right)$ 。

虽然随机投影具有良好的性质，但在应用随机投影解决大规模机器学习任务时仍然面临一些问题。

- 虽然随机矩阵的产生非常容易，但当数据量 n 和原维度 d 都非常大时，(1)中的投影计算复杂度会变得非常高，难以对所有样本进行降维。
- 在早期的研究工作中，随机投影被认为是一种数据预处理的方式，并没有和学习任务结合起来，因此随机投影对学习任务尤其是监督学习会造成什么样的影响并不清楚。

针对第一个问题，研究人员提出了一系列快速随机投影算法，主要包括稀疏随机投影^[40-42]以及具有特殊结构的快速投影^[43,44]。前者通过生成一个稀疏的投影矩阵，使得投影操作只跟矩阵的非零元素有关。比如 Ailon 和 Chazelle^[40]提出的稀疏投影算法能够将单个样本的投影计算复杂度从 $O(md)$ 降低到 $O(d \log d + m^3)$ 。后者通过生成具有特殊结构的投影矩阵，加速投影操作。比如 Liberty 等人^[44]提出的 Lean Walsh 矩阵能够将单个样本的投影计算复杂度降低到 $O(d)$ 。

针对第二个问题，越来越多的研究人员开始将随机投影与学习任务联系起来，从理论上分析随机投影对学习问题的影响。Balcan 等人^[45]和 Shi 等人^[46]分别研究了随机投影对分类问题的影响，他们的结论表明，如果数据在原空间内存在一个大的分类间隔，那么随机投影后的数据仍然是近似可分的。Drineas 等人^[47]将随机投影应用到求解最小二乘问题中，利用随机投影，最小二乘的计算复杂度由 $O(nd^2)$ 降低到了 $O(nd \log d)$ 。Paul 等人^[48]研究了随机投影对支持向量机的影响，结果表明当数据矩阵满足低秩假设时，随机投影能够保持良好的泛化能力。

近年来，Zhang 等人^[49,50]和 Yang 等人^[51]提出从优化的角度分析随机投影造成的影响。Zhang 等人^[49,50]首先证明了数据降维后得到的分类器与原空间内的分类器存在很大的距离，因此无法用来指导原空间内的特征选择。基于此，他们提出了一种对偶随机投影算法，通过对偶问题作为桥梁恢复原空间内的分类器。理论分析表明，当数据矩阵的秩为 r 时，只需要将数据投影到 $\Omega(r \log r)$ - 维空间内就可以近似恢复出原空间内的分类器。当数据不满足低秩假设时，Yang 等人^[51]提出了一种对偶问题稀疏性的随机投影算法，利用 ℓ_1 - 正则化技术得到对偶问题的准确解。

2.2 算法伸缩

2.2.1 分布式学习

分布式机器学习（简称分布式学习）采用“分而治之”的思想，将一个大规模任务分解成多个小规模子任务，然后让集群中的多台机器各自处理其中的一个或者多个子任务，最后对子任务的中间计算结果进行归约得到原始大规模任务的解。因此，分布式学习通过将计算分摊到多台机器并行执行，可以加速学习过程。例如：在理想情况下， k 台机器并行完成一个任务的时间只需要单台机器的 $1/k$ 。因此，分布式学习可以加强计算

的可扩展性，在处理超大规模问题时展现了巨大的潜能，已经成为大数据机器学习的研究热点之一，引起了学术界和工业界的高度关注。分布式学习方面的研究既包括分布式编程模型等系统架构方面的工作，比如 HadoopMapReduce^[52]、Spark^[53]、GraphLab^[54,55]、Parameter Server (PS)^[56-60]等，还包括分布式学习算法的设计，比如交替方向乘子法 (ADMM)^[61]等。研究难点包括如何对学习任务进行分解、如何实现不同机器之间的负载均衡，以及如何降低机器之间的通信开销。

在分布式编程模型等系统架构方面，近年来出现了多种面向大数据机器学习的解决方案，它们都有各自的问题解决思路，旨在解决多种多样的大规模机器学习问题。Hadoop MapReduce^[52]通过将任务分解成 Map 和 Reduce 两个阶段，可以实现分而治之的思想。在 Map 阶段，整个任务被划分成多个子任务并分配给不同的机器分别执行，然后在 Reduce 阶段实现对多个子任务的归约。Hadoop MapReduce 虽然能很好地完成某些简单的计算任务，例如计算矩阵与向量的乘积，但不适合于处理复杂机器学习任务。主要原因是复杂机器学习任务通常需要多轮循环迭代，而 MapReduce 在每轮迭代都需要从硬盘文件读写数据，从而导致效率非常低下。Spark^[53]提出了内存计算的思想，解决了 Hadoop MapReduce 中不断进行文件读写的开销问题。因此，相对于 Hadoop MapReduce，Spark 更适合于需要多轮迭代计算的任务。但 Spark 要求将模型转换成 RDD 的模式，在有些问题上并不是很高效，而且在通信优化方面考虑得也不多。另外，Spark 并不是为大数据机器学习专门设计的，因此，目前提供的机器学习接口还不是很多，而且也没有进行相应的优化，性能不是特别理想。

GraphLab^[54,55]是为了解决基于图的机器学习而提出的，采用了“像点一样思考”(think like a vertex)的思想，将图中每个顶点作为独立的并行计算单位，执行用户定义的点中心程序 (vertex-centric programs)，并在边上完成通信来实现整个图的分布式计算。分布式图计算（学习）中影响性能的最关键因素是图的划分，即将图的顶点和边分配给集群中的不同机器。图划分算法的好坏直接影响整个系统的负载均衡和通信开销。最初版本的 GraphLab^[54] (GraphLab1) 采用切割边的方式将顶点分配到不同的机器，后来的改进版本 (GraphLab2，即 PowerGraph)^[55] 采用切割顶点的方式将边分配到不同的机器。理论和实验都证明，采用切割顶点的图划分方式在服从幂律分布的图上能得到比切割边的图划分方式更好的性能。因此，现在我们所说的 GraphLab 一般指的是 GraphLab2，即 PowerGraph^[55]。GraphLab 为基于图的机器学习提供了良好的编程接口，但对于通用的机器学习任务处理比较困难。

Parameter Server (PS)^[56-60]，即参数服务器，是专门针对大规模分布式机器学习提出的一种编程模型。在 PS 中，模型的参数被分布式地存储在一组被称为参数服务器的机器节点中，而训练数据被分布式地存储在另一组被称为计算服务器的机器节点中。一个计算节点可以从参数服务器上获取模型的最新参数值，也可以把自己计算得到的结果发送到所有参数服务器节点。不同的计算节点并行地执行分配给自己的子任务。PS 是一种通用的架构，有不同的研究人员或者机构实现了不同的 PS 版本，例如，Petuum^[56,57] 和 Mu Li 等人开发的 PS^[58,59]都是开源的，而 Google 内部用于训练深度神经网络的 PS^[60]是不开

源的。不同 PS 版本在解决通信开销等系统实现细节上存在着差异，因此，在处理不同任务时表现出的性能可能也不一样。PS 能实现包括基于图的机器学习在内的大部分通用机器学习任务。

虽然上述分布式编程模型会在系统容错和通信开销等方面进行一定的优化，使得用户编程更加简便和透明，但也有部分研究者直接基于传统的 MPI (message passing interface) 实现分布式机器学习，例如，Chapelle 等人^[62] 直接基于 MPI 实现分布式 Logistic Regression，并成功应用于企业级的广告点击率预估问题。

在分布式学习算法设计方面，目前很多算法都基于交替方向乘子法 (ADMM)^[61] 来实现。ADMM 为分布式学习提供了一种很自然的形式化方法。假设集群中共有 P 台机器，训练数据被相应地划分为 P 份，每台机器上被分配到一份。 \mathbf{w} 是需要学习的参数， $f_i(\mathbf{w})$ 表示第 i 台机器上的数据导致的损失函数值，那么，所有训练数据导致的损失函数值为：

$$f(\mathbf{w}) = \sum_{i=1}^P f_i(\mathbf{w}) \quad (3)$$

最小化上述损失函数等价于如下形式：

$$\begin{aligned} & \min \sum_{i=1}^P f_i(\mathbf{w}_i) \\ & \text{s. t. } \mathbf{w}_i - \mathbf{w} = \mathbf{0}, i = 1, \dots, P \end{aligned} \quad (4)$$

其中， \mathbf{w}_i 是第 i 台机器上的局部变量， \mathbf{w} 是全局变量。通常我们把保存局部变量的机器叫作从节点，把保存全局变量的机器叫作主节点。通过这样变换后，就可以利用 ADMM 框架，方便地实现整个损失函数的分布式学习。也就是说，每个从节点只基于分配给它的部分数据，更新局部的 \mathbf{w}_i ，并把更新后的值发送给主节点，等主节点收集到所有从节点的信息后，主节点更新全局变量 \mathbf{w} ，然后再把 \mathbf{w} 发送给每个从节点，从节点收到主节点的信息后进一步更新对应的局部变量，循环迭代直到满足停止条件。

上述过程实际上是采用了同步机制，即更新全局变量的时候，得等所有局部变量更新完才能进行。如果集群中不同机器的负载不平衡或者机器性能不一样，某台慢的机器会造成整个系统在等待上浪费时间，从而造成系统性能的下降。最近有研究者提出了半同步的 ADMM 算法^[63] 来解决这个问题。基本思路是主节点在收集到部分从节点的信息后，就可以更新全局变量，不需要等待最慢的那些从节点。而当从节点的更新轮数比主节点的更新轮数小（慢）到一定阈值后，直接从主节点获取最新的全局变量，从而赶上整个系统的进度。理论证明该策略能保证算法的收敛性，同时也能减少通信开销，提高整个系统的效率。

最近，也有研究者针对下一节将要介绍的随机梯度下降 (stochastic gradient descent, SGD) 方法提出了一系列分布式实现策略^[64-68]。

2.2.2 随机优化

许多机器学习问题比如支持向量机，在经过数学建模后都会转换为优化问题。如何

快速求解这类优化问题，是大数据机器学习的一项重要研究内容。随机优化（stochastic optimization）^[69]泛指一类只需要获取目标函数随机无偏梯度的优化技术，被广泛应用于求解机器学习问题。下面我们首先介绍通用的随机优化技术，然后介绍针对传统随机优化技术不足而提出的改进，重点关注目标函数为有限项之和的最新研究进展。

随机优化考虑如下问题

$$\min_{\mathbf{w} \in \mathcal{W}} f(\mathbf{w}) \quad (5)$$

其中 $f(\cdot)$ 为目标函数， \mathcal{W} 是定义域。为了便于分析，通常假设 $f(\cdot)$ 是凸函数， \mathcal{W} 是凸集合。对于随机优化，给定任意一个解 $\mathbf{w} \in \mathcal{W}$ ，我们假设能够快速地求解函数 $f(\cdot)$ 在该点上的无偏随机梯度 $g(\mathbf{w})$ ，即

$$E[g(\mathbf{w})] = \nabla f(\mathbf{w})$$

求解随机优化的典型算法为随机梯度下降（stochastic gradient descent, SGD）^[70]，其主要流程如下：

```

for  $t = 1, \dots, T$  do
     $\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t g(\mathbf{w}_t)$ 
     $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$ 
end for

```

其中， \mathbf{w}_t 代表第 t 次迭代的初始解， $\eta_t > 0$ 为步长， $\Pi_{\mathcal{W}}(\mathbf{x}) = \arg \min_{\mathbf{w} \in \mathcal{W}} \|\mathbf{w} - \mathbf{x}\|_2^2$ 为集合 \mathcal{W} 上的投影操作。在每一次迭代，首先沿着随机梯度 $g(\mathbf{w}_t)$ 的反方向更新 \mathbf{w}_t ，得到临时解 \mathbf{w}'_{t+1} ，这也是SGD名称的由来；然后，为了保证最终解属于 \mathcal{W} ，需要对 \mathbf{w}'_{t+1} 进行投影操作以得到下一次迭代的解 \mathbf{w}_{t+1} 。一般情况下，我们将 \mathbf{w}_1 设置为 \mathcal{W} 中的任意点，在算法结束后用平均解 $\frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ 作为最后输出。

根据在线梯度下降的理论分析^[71,72]以及随机变量的集中不等式^[73]，可以证明在随机梯度和定义域有界的前提下，对于利普希茨连续凸函数，SGD返回的平均解可以取得 $O(1/\sqrt{T})$ 的收敛速率；对于强凸函数，SGD返回的平均解可以达到 $O(\log T/T)$ 的收敛速率。理论分析表明^[74,75]，对于利普希茨连续凸函数而言， $O(1/\sqrt{T})$ 已经是最优的收敛速率，但对于强凸函数，其最优速率是 $O(1/T)$ 。因此，传统的SGD对于强凸函数并非最优的算法。自2010年以来，研究人员提出了一些SGD的变形^[76-78]，对于强凸函数能够取得最优的 $O(1/T)$ 的收敛速率，其中最著名的算法包括Hazan和Kale^[77]提出的Epoch Gradient Descent (Epoch-GD) 和 Rakhlin等人^[78]提出的SGD with α -suffix Averaging (SGD $_{\alpha}$)。Epoch-GD在传统的SGD算法上加了一层外循环，在算法运行过程中，每一轮外循环采用上一轮外循环返回的值作为初始解。SGD $_{\alpha}$ 更加简单，只对SGD后 αT 个解 $\mathbf{w}_{(1-\alpha T)+1}, \dots, \mathbf{w}_T$ 求平均作为最后输出。需要注意的是，这里介绍的几种算法都需要对SGD全部或部分中间解求平均。在2013年，Shamir和Zhang^[79]分析了SGD最后一次迭代 \mathbf{w}_T 的收敛速率。其结果表明对于非平滑凸函数， \mathbf{w}_T 取得了 $O(\log T/\sqrt{T})$ 的收敛速率；

对于非平滑强凸函数， \mathbf{w}_T 取得了 $O(\log T/T)$ 的收敛速率。

由于随机梯度可以很容易计算，SGD 非常适用于大数据机器学习，尤其是大规模监督学习问题。给定一组训练数据 $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ ，其中 $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \mathbf{R}$ 。根据经验误差最小化原则，监督学习通过求解下面的优化问题得到一个参数为 $\mathbf{w} \in \mathbf{R}^d$ 的线性分类器：

$$\min_{\mathbf{w} \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n \ell(y_i \mathbf{x}_i^\top \mathbf{w}) + \lambda \Phi(\mathbf{w}) \quad (6)$$

其中 $\ell(\cdot) : \mathbf{R} \mapsto \mathbf{R}$ 是用来衡量拟合误差的损失函数，如 SVM 中的铰链损失： $\ell(u) = \max(0, 1 - u)$ ， $\Phi(\cdot)$ 是用来防止过拟合的正则化函数，如 ℓ_2 -范数平方： $\Phi(\mathbf{w}) = \|\mathbf{w}\|_2^2$ ， $\lambda > 0$ 是正则化参数。如果我们采用传统的梯度下降算法，每一次迭代都需要计算(6)中目标函数的梯度，其计算复杂度为 $O(nd)$ 。对于大数据而言， n 的数值非常大，因此无法直接采用梯度下降法求解该问题。采用 SGD，在每一轮迭代，我们只需要随机选择一个训练样本，然后利用该样本更新当前的模型。假设当前为第 t 次迭代，SGD 首先随机生成一个索引 $i_t \in \{1, 2, \dots, n\}$ ，然后仅仅利用 $(\mathbf{x}_{i_t}, y_{i_t})$ 更新当前解 \mathbf{w}_t ：

$$\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t [\nabla \ell(y_{i_t} \mathbf{x}_{i_t}^\top \mathbf{w}_t) + \lambda \nabla \Phi(\mathbf{w}_t)] \quad (7)$$

上式的计算复杂度为 $O(d)$ ，与样本数 n 无关，因此适用于大数据场景。

尽管 SGD 在大规模机器学习上取得了巨大的成功，但传统的 SGD 仍然存在一些不足，集中表现在下面两方面：

- 传统的 SGD 忽略了投影操作 $\Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$ 的计算复杂度。当我们面临复杂的定义域时，比如半正定锥，投影操作的计算复杂度会变得非常高，成为计算瓶颈。
- 与梯度下降相比，SGD 的收敛速率过慢，难以取得精确解。当目标函数是平滑凸函数时，梯度下降能达到 $O(1/T^2)$ 的收敛速率^[80,81]，而 SGD 只有 $O(1/\sqrt{T})$ 的收敛速率；同理，对于平滑并且强凸函数，梯度下降可以达到指数收敛，即 $O(1/\alpha^T)$ （其中 $\alpha > 1$ ）^[80,82]，而 SGD 只能达到 $O(1/T)$ 的收敛速率。

针对第一个问题，研究人员提出各种手段来减少投影的次数。当目标函数是平滑函数时，可以通过小批量（Mini-batch）的技术来减少投影的次数。在每一次迭代过程中，小批量技术利用多个随机梯度的均值来更新当前解。虽然更新的次数变少了，但由于平均梯度更加准确，收敛速率可以保持不变。对于平滑凸函数，Cotter 等人^[83] 证明了利用小批量可以把投影的次数从 $O(T)$ 降低到 $O(T^{1/4})$ ，而收敛速率依旧保持在 $O(1/\sqrt{T})$ ，这里 T 代表随机梯度的个数。对于平滑并且强凸函数，Zhang 等人^[84] 证明了利用小批量可以将投影次数从 $O(T)$ 降低到 $O(\log T)$ ，同时保持 $O(1/T)$ 的收敛速率。此外，对于某些特殊的定义域 \mathcal{W} ，比如半正定锥，我们可以采用 Frank-Wolfe 方法^[85,86] 来避免投影操作。其核心思想是将投影操作转换为一个 \mathcal{W} 上的线性优化问题。最后，对于定义域为等号线性约束的情况，我们可以采用随机交替方向乘子法（Stochastic ADMM）^[87,88] 来避免投影操作，其核心思想是将约束条件通过拉格朗日函数的形式引入到目标函数中。

针对第二个问题，近年来研究人员提出了一系列算法来利用(6)中目标函数的特殊结构来提升 SGD 的收敛速率，成为机器学习领域的研究热点^[89-98]。这里的结构是指(6)中的目标函数可以写成 n 个函数求和的形式。其中最著名的算法包括 Roux 等人提出的

Stochastic Average Gradient (SAG)^[89], Shalev-Shwartz 和 Zhang 提出的 Stochastic Dual Coordinate ascent (SDCA)^[90], Zhang 等人提出的 Epoch Mixed Gradient Descent (EMGD)^[92], 以及 Johnson 和 Zhang 提出的 Stochastic Variance Reduced Gradient (SVRG)^[94]。当目标函数平滑并且强凸时, 这些算法采用不同的策略控制随机梯度的方差, 都取得了线性收敛速率。SAG、EMGD 和 SVRG 直接优化(6)中的目标函数, 而 SDCA 则是优化其对偶问题。其中 EMGD 和 SVRG 两种算法几乎完全类似, 主要区别在于理论分析及问题设定。

下面我们给出 EMGD/SVRG 的算法流程。为了表示方便, 记 $f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(y_i \mathbf{x}_i^\top \mathbf{w}) + \lambda \Phi(\mathbf{w})$, $f_i(\mathbf{w}) = \ell(y_i \mathbf{x}_i^\top \mathbf{w}) + \lambda \Phi(\mathbf{w})$, 显然 $f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{w})$ 。

```

for  $k = 1, \dots, m$  do
    Set  $\mathbf{w}_1 = \tilde{\mathbf{w}}_k$ 
    Let  $\mathbf{g} = \nabla f(\mathbf{w}_1)$ 
    for  $t = 1, \dots, T$  do
        Random Sample  $i_t \in [n]$ 
         $\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t (\nabla f_{i_t}(\mathbf{w}_t) - \nabla f_{i_t}(\mathbf{w}_1) + \mathbf{g})$ 
         $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$ 
    end for
    Se  $\tilde{\mathbf{w}}^{k+1} = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$ 
end for

```

与 Epoch-GD^[77]类似, EMGD/SVRG 也包含两重循环。在第 k 次外循环, 我们将上次外循环得到的解 $\tilde{\mathbf{w}}_k$ 作为内循环的初始值; 然后计算该初始值的真实梯度 $\mathbf{g} = \nabla f(\mathbf{w}_1)$; 接下来在每一轮内循环, 我们随机选择一个索引 i_t , 并利用混合梯度 $\nabla f_{i_t}(\mathbf{w}_t) - \nabla f_{i_t}(\mathbf{w}_1) + \mathbf{g}$ 更新内循环 \mathbf{w}_t 。可以验证, 这里的混合梯度依然是 \mathbf{w}_t 真实梯度 $\nabla f(\mathbf{w}_t)$ 的无偏估计。利用混合梯度的优点在于当函数平滑时, 混合梯度的方差可以被 \mathbf{w}_t 和 \mathbf{w}_1 之间的距离控制住。理论分析表明, 对于无约束情况, EMGD/SVRG 通过使用 $O(\log 1/\epsilon)$ 个真实梯度以及 $O(\kappa \log 1/\epsilon)$ 个随机梯度, 就可以找到一个 ϵ -最优解, 其中 κ 为优化问题的条件数 (condition number)^[94]。

3 国内研究进展

在哈希学习方面, 国内学者已进行了有益的探索。香港科技大学 Dit-Yan Yeung、杨强等人分别在多模态哈希学习、异质数据哈希学习方面进行了研究^[99-101]。南京大学

LAMDA 研究组的李武军等人对非监督哈希学习、监督哈希学习、多模态哈希学习、按位优化（学习）策略、量化策略等方面进行了研究^[102-107]。浙江大学何晓飞、蔡登等人研究了非监督和半监督哈希学习方法^[108,109]。该方面技术在信息检索、计算机视觉和多媒体领域得到了广泛应用，例如复旦大学黄萱菁等人将哈希学习用于内容重用检测^[110]；中科院计算所陈熙霖、山世光研究组研究了多模态哈希学习和半监督哈希学习并用于跨模态相似度检索和图像检索^[111,112]；中科院自动化所刘成林研究组将哈希用于实现快速的图构造^[113]，卢汉清、徐常胜、程健等人研究了不同的哈希学习方法并用于多媒体和视觉对象检索^[114,116]；清华大学朱文武、杨士强等人研究了异构哈希学习算法并用于跨模态检索^[117]，丁贵广等人研究了多模态与跨模态相似度检索中的哈希学习方法^[118,119]；浙江大学庄越挺和吴飞等人研究了稀疏的多模态哈希学习方法^[120]；西安电子科技大学高新波等人研究了多模态哈希学习并用于视频拷贝检测等应用^[121,122]；西安光机所李学龙等人研究了基于局部模型的鲁棒哈希学习和谱嵌入哈希学习并用于图像检索应用^[123,124]。此外，中山大学郑伟诗等人研究了在线哈希学习方法^[125]，电子科技大学沈复民等人研究了流形（图）上的归纳式（inductive）哈希学习方法^[126]。

在随机投影方面，南京大学 LAMDA 研究组的张利军等人先后提出了对偶随机投影算法^[49,50]和利用对偶稀疏性的随机约减算法^[51]，利用随机投影加速优化问题的求解；浙江大学何晓飞等人^[127]提出了一种基于随机投影的近似重复查询算法；北京大学查红彬等人^[128]提出了一种基于随机投影的鲁棒纹理分类算法。

在分布式学习方面，国内的企业界和学术界都已经开展了一定的探索。国内很多企业，尤其是互联网企业，都已经广泛采用相关平台和系统（例如 Spark 等）来实现大数据的分析和挖掘。在分布式编程模型等系统架构方面，香港科技大学杨强研究组提出了一种针对大规模时空移动宽带数据的分布式分析系统^[129]；上海交通大学陈海波研究组提出了一种分布式图计算框架 PowerLyra^[130]，通过结合切割顶点和切割边两种图划分方法，改善了 PowerGraph 的性能；南京大学李武军、上海交通大学张志华等人提出了一种新的切割顶点的图划分方法^[131]，并从理论和实验上证明了该方法在幂律图上取得了比已有方法（PowerGraph）更好的性能。在分布式学习算法设计上，香港科技大学郭天佑等人提出了半同步的 ADMM 分布式学习算法^[63]；清华大学张钹、朱军等人提出了一种分布式贝叶斯后验概率采样方法^[132,133]；上海交通大学张志华等人提出了一种分布式 Fisher 判别分析方法^[134]；南京大学 LAMDA 研究组的李武军等人基于 ADMM 设计了一种分布式随机梯度下降算法，用于实现大规模的矩阵分解^[135]，并基于 MPI 实现了一种互联网广告点击率预估模型的大规模分布式学习^[136]。

在随机优化方面，香港科技大学郭天佑等人提出了针对随机优化的加速梯度法^[137]和快速随机交替方向乘子法^[138]；南京大学 LAMDA 研究组进行了较多探索，例如提出了基于小批量的随机优化算法^[84]、达到线性收敛速率的混合梯度下降^[92]、大规模随机交替方向乘子法^[88]等；清华大学张长水研究组^[139]将随机优化的思路应用到在线核学习中，提出了一种快速的非线性学习算法；浙江大学何晓飞^[140]等人提出了一种针对复合函数的稀疏随机优化算法。

4 发展趋势与展望

在哈希学习方面，目前大部分哈希学习研究的思路为：针对某个机器学习场景（比如排序学习场景^[22]）或者应用场景，只要以前没有人尝试过用哈希学习的思想来加速学习过程，就可以考虑把哈希学习用进去，然后在一个传统模型（这个传统模型不用哈希学习）解决不了的数据或者应用规模上进行实验验证。从解决实际问题的角度来讲，这些工作虽然初步，但还是很有研究价值的，毕竟为大数据中传统模型不能解决的问题提供了一种可行的解决思路。但从哈希学习本身的研究来讲，目前大部分工作还没有从哈希学习问题的本质上进行考虑。因此，哈希学习虽已被广泛关注并在某些应用领域取得了初步成效，但研究才刚刚开始，问题本质和模型构建有待于进一步深入思考，模型参数的优化方法有待于进一步探索，量化阶段的重要性已经引起注意，但量化策略期待进一步突破。另外，大部分学习场景和应用领域到目前为止还只出现很少的哈希学习方法，有的场景和应用甚至还没有研究者进行哈希学习的尝试。例如，推荐系统是个很大的应用方向，但到目前为止这方面采用哈希学习的工作还不多^[28]。因此，怎样将哈希学习的思想和方法拓展到新的学习场景和应用领域，用来解决传统方法在遇到大数据时不能解决的问题，将是非常有意义的工作。特别值得一提的是，很多分布式机器学习的瓶颈在于节点间的通信开销。因此，将哈希学习引入到分布式机器学习算法，并验证哈希学习在减小通信开销方面的有效性，也是非常有意义的研究方向。

在随机投影方面，将随机投影与具体的机器学习任务结合起来是目前的研究热点。但现有的理论成果仍然局限于最小二乘、支持向量机、逻辑回归这些简单的监督学习问题中。对于其他类型的监督学习，比如稀疏学习、深度学习，随机投影缺乏算法和理论支撑。因此，进一步完善和丰富基于随机投影的机器学习算法和理论是一个重要的研究方向。此外，虽然基于随机投影的学习算法能够大幅度降低计算复杂度，但其不可避免地引入了误差。如何进一步降低随机投影算法的误差是一个尚未解决的问题。在今年的国际机器学习大会（ICML 2015）上，Yang 等人^[51]提出将随机投影得到的解作为初始值传递给其他优化算法，这样不仅能降低总体计算时间还可以得到非常精确的解，为解决该问题提供了一种可能的思路。

在分布式学习方面，目前已经出现了多种分布式编程模型，这为推动大数据机器学习的快速发展提供了良好的条件和基础。但现在面临的一个重要问题是，还没有工作对这些不同的编程模型进行详细对比和基准测试。因此，研究人员或者用户在面临具体问题的时候，并不清楚该选择何种编程模型更合适。对已有编程模型进行定性和定量两方面详细的对比和测试，找出它们之间的优劣，将是一项非常有意义的工作。另外，分布式编程模型方面的研究才刚刚开始，里面有大量的问题值得研究，包括系统、算法、理论和应用等。在分布式学习算法设计方面，现在很多研究者很自然地想到利用 ADMM，但 ADMM 是否就是最好的算法框架，还值得进一步深入研究。

在随机优化方面，提升随机优化的收敛速率是一个重要的研究方向，但目前该方面的工作都需要假设目标函数为平滑并强凸。在实际问题中，我们遇到的目标函数未必满足该性质，比如铰链损失不平滑、逻辑损失不强凸。如何提升随机优化处理非平滑或非强凸函数的收敛速率^[141]，是一个亟待解决的问题。此外，目前针对随机优化的理论研究主要集中在凸优化问题上，但在实践中，随机优化（主要是 SGD 算法）已经被成功应用于求解非凸问题，比如深度学习的训练中^[60,142]。因此，从理论上迫切需要建立非凸随机优化的性能保证。在今年召开的学习理论会议上（COLT 2015）上，Ge 等人^[143]证明了对于某类特殊的非凸优化问题，如张量分解，SGD 能够在多项式时间内收敛到局部最优解。这一工作拓展了我们对 SGD 收敛性能的理解，对分析非凸随机优化有重要的借鉴意义。

5 结束语

从 2015 年国际机器学习大会（ICML）的几个热门关键词：“深度学习”、“优化”、“在线学习”可看出，机器学习领域非常关注大数据的“巨量”、“多样”和“快速”等特性，本报告简介了近几年出现的、为解决数据的巨量性而提出的新型机器学习模型与算法。值得强调的是，大数据机器学习已成为人工智能领域最受关注的研究内容之一，并将被广泛应用于其他学科和领域，从而产生重大和深远的影响。

致谢

本报告受中国计算机学会人工智能与模式识别专委会委托撰写；感谢中国计算机学会学术工委对选题的建议。报告中关于哈希学习的部分主要来自^[144]。因时间所限仓促成稿，报告中对相关进展的介绍难免挂一漏万，仅供参考。

参考文献

- [1] Viktor Mayer-Schönberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think [M]. Eamon Dolan / Houghton Mifflin Harcourt, 2013.
- [2] T Hey, S Tansley, K Tolle. The Fourth Paradigm: Data-Intensive Scientific Discovery [M]. Microsoft Research, 2009.
- [3] 周志华. 大数据对机器学习的挑战 [R]. 未来数据论坛，主题报告，2013.
- [4] Z-H Zhou, N V Chawla, Y Jin, et al. Big data opportunities and challenges: Discussions from data analytics perspectives [J]. IEEE Computational Intelligence Magazine, 2014, 9(4) : 62-74.
- [5] M I Jordan. Message from the President: The era of Big Data [S]. ISBA Bulletin, 18(2) : 1-3, 2011.

- [6] CCF人工智能与模式识别专业委员会. 深度学习的研究进展与趋势[R]. CCF2013—2014 中国计算机科学技术发展报告, 机械工业出版社, 2014 : 119-138.
- [7] R Salakhutdinov, G Hinton. Semantic Hashing[R]. In SIGIR Workshop on Information Retrieval and Applications of Graphical Models , 2007.
- [8] R Salakhutdinov, G Hinton. Semantic hashing[J]. International Journal of Approximate Reasoning , 2009 , 50(7) : 969-978.
- [9] Y Weiss, A Torralba, R Fergus. Spectral hashing[C]. In Advances in Neural Information Processing Systems 21 (NIPS), 2008.
- [10] B Kulis, P Jain, K Grauman. Fast similarity search for learned metrics[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) , 2009 , 31(12) : 2143-2157.
- [11] K Weinberger, A Dasgupta, J Langford, et al. Feature hashing for large scale multitask learning[C]. In Proceedings of the 26th International Conference on Machine Learning (ICML) , 2009.
- [12] J Wang, S Kumar, S-F Chang. Semi-supervised hashing for large-scale search[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) , 2012 , 34(12) : 2393-2406.
- [13] Y Gong, S Lazebnik, A Gordo, et al. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) , 2013 , 35(12) : 2916-2929.
- [14] W Liu, J Wang, S Kumar, et al. Chang. Hashing with graphs[C]. In Proceedings of the 28th International Conference on Machine Learning (ICML) , 2011.
- [15] M Rastegari, J Choi, S Fakhrاء, et al. Predictable dual-view hashing[C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [16] A Gionis, P Indyk, R Motwani. Similarity search in high dimensions via hashing[C]. In Proceedings of the 25th International Conference on Very Large Data Bases (VLDB) , 1999.
- [17] M Datar, N Immorlica, P Indyk, et al. Locality-sensitive hashing scheme based on p-stable distributions [C]. In Proceedings of the 20th ACM Symposium on Computational Geometry (SOCG) , 2004.
- [18] A Andoni, P Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [J]. Communication of ACM (CACM) , 2008 , 51(1) : 117-122.
- [19] M Norouzi, D J Fleet. Minimal loss hashing for compact binary codes[C]. In Proceedings of the 28th International Conference on Machine Learning (ICML) , 2011.
- [20] M Norouzi, D J Fleet, R Salakhutdinov. Hamming distance metric learning[C]. In Advances in Neural Information Processing Systems 25 (NIPS) , 2012.
- [21] C Strecha, A M Bronstein, M M Bronstein, et al. LDAhash: Improved matching with smaller descriptors [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) , 2012 , 34(1) : 66-78.
- [22] X Li, G Lin, C Shen, et al. Learning hash functions using column generation[C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [23] D Zhang, F Wang, L Si. Composite hashing with multiple information sources[C]. In Proceedings of the 34th ACM Conference on Research and Development in Information Retrieval (SIGIR) , 2011.
- [24] J He, W Liu, S-F Chang. Scalable similarity search with optimized kernel hashing[C]. In Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) , 2010.
- [25] W Liu, J Wang, R Ji, et al. Supervised hashing with kernels[C]. In Proceedings of the 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2012.

- [26] X Zhu, Z Huang, H T Shen, et al. Linear cross-modal hashing for efficient multimedia search [C]. In Proceedings of the 21st ACM international conference on Multimedia (MM), 2013.
- [27] H Xu, J Wang, Z Li, et al. Complementary hashing for approximate nearest neighbor search [C]. In Proceedings of the 13rd IEEE International Conference on Computer Vision (ICCV), 2011.
- [28] K Zhou, H Zha. Learning binary codes for collaborative filtering [C]. In Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2012.
- [29] P Sarkar, D Chakrabarti, M I Jordan. Nonparametric link prediction in dynamic networks [C]. In Proceedings of the 29th International Conference on Machine Learning (ICML), 2012.
- [30] A Bellet, A Habrard, M Sebban. A survey on metric learning for feature vectors and structured data [J/OL]. arXiv: 1306.6709, 2013.
- [31] S Moran, V Lavrenko, M Osborne. Variable bit quantization for LSH [C]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), 2013.
- [32] S S Vempala. The Random Projection Method [M]. American Mathematical Society, 2004.
- [33] S Dasgupta. Experiments with random projection [C]. In Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI), 2000.
- [34] E Bingham, H Mannila. Random projection in dimensionality reduction: applications to image and text data [C]. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD), 2001.
- [35] C M Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics) [M]. Springer, 2007.
- [36] S Dasgupta, A Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss [J]. Random Structures & Algorithms, 2003, 22(1): 60-65.
- [37] D Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins [J]. Journal of Computer and System Sciences, 2003, 66(4): 671-687.
- [38] R I Arriaga, S Vempala. An algorithmic theory of learning: robust concepts and random projection [J]. Machine Learning, 2006, 63(2): 161-182.
- [39] A Magen. Dimensionality reductions that preserve volumes and distance to affine spaces, and their algorithmic applications [R]. Randomization and Approximation Techniques in Computer Science, Lecture Notes in Computer Science 2483, 2002.
- [40] N Ailon, B Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform [C]. In Proceedings of the 38th annual ACM symposium on Theory of computing (STOC), 2006.
- [41] A Dasgupta, R Kumar, T Sarlos. A sparse Johnson-Lindenstrauss transform [C]. In Proceedings of the 42nd ACM symposium on Theory of computing (STOC), 2010.
- [42] D M Kane, J Nelson. Sparser Johnson-Lindenstrauss transforms [C]. In Proceedings of the 23rd annual ACM-SIAM symposium on Discrete Algorithms (SODA), 2012.
- [43] N Ailon, E Liberty. Fast Dimension Reduction Using Rademacher Series on Dual BCH Codes [J]. Discrete & Computational Geometry (DCG), 2009, 42(4): 615-630.
- [44] E Liberty, N Ailon, A Singer. Dense fast random projections and Lean Walsh transforms [J]. Discrete & Computational Geometry (DCG), 2011, 45(1): 34-44.
- [45] M F Balcan, A Blum, S Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings [J]. Machine Learning, 2006, 65(1): 79-94.

- [46] Q Shi, C Shen, R Hill, et al. Hengel. Is margin preserved after random projection[C]. In Proceedings of the 29th International Conference on Machine Learning (ICML) , 2012.
- [47] P Drineas, M W Mahoney, S Muthukrishnan, et al. Faster Least Squares Approximation[J]. Numerische Mathematik , 2011 , 117(2) : 217-249.
- [48] S Paul, C Boutsidis, M Magdon- Ismail, et al. Random projections for support vector machines[C]. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS) , 2013.
- [49] L Zhang, M Mahdavi, R Jin, et al. Recovering the Optimal Solution by Dual Random Projection[C]. In Proceedings of the 26th Annual Conference on Learning Theory (COLT) , 2013.
- [50] L Zhang, M Mahdavi, R Jin, et al. Random Projections for Classification: A Recovery Approach[J]. IEEE Transactions on Information Theory (TIT) , 2014 , 60(11) : 7300-7316.
- [51] T Yang, L Zhang, R Jin, et al. Theory of Dual- sparse Regularized Randomized Reduction [C]. In Proceedings of the 32nd International Conference on Machine Learning (ICML) , 2015.
- [52] Tom White. Hadoop: The Definitive Guide[M]. 1st ed. O'Reilly Media. 2009.
- [53] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets [C]. In Proceedings of the 2nd USENIX conference on Hot topics in cloud computing (HotCloud) , 2010.
- [54] Y Low, J Gonzalez, A Kyrola, et al. GraphLab. A New Framework for Parallel Machine Learning[C]. In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI) , 2010.
- [55] J E Gonzalez, Y Low, H Gu, et al. PowerGraph: distributed graph-parallel computation on natural graphs [C]. In Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation (OSDI) , 2012.
- [56] Q Ho, J Cipar, H Cui, et al. More effective distributed ML via a stale synchronous parallel parameter server[C]. In Advances in Neural Information Processing Systems 26 (NIPS) , 2013.
- [57] E P Xing, Q Ho, W Dai, et al. Petuum: a new platform for distributed machine learning on big data[C]. In Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) , 2015.
- [58] M Li, D G Andersen, J W Park, et al. Scaling distributed machine learning with the parameter server [C]. In Proceedings of the 11th USENIX conference on Operating Systems Design and Implementation (OSDI) , 2014.
- [59] M Li, D G Andersen, A J Smola, et al. Communication efficient distributed machine learning with the parameter server[C]. In Advances in Neural Information Processing Systems 27 (NIPS) , 2014.
- [60] J Dean, G S Corrado, R Monga, et al. Large scale distributed deep networks[C]. In Advances in Neural Information Processing Systems 25 (NIPS) , 2012.
- [61] S Boyd, N Parikh, E Chu, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends in Machine Learning , 2011 , 3(1) : 1-122.
- [62] O Chapelle, E Manavoglu, R Rosales. Simple and scalable response prediction for display advertising[J]. ACM Transactions on Intelligent Systems and Technology (TIST) , 2015 , 5(4) : 1-34.
- [63] R Zhang, J T Kwok. Asynchronous distributed ADMM for consensus optimization[C]. In Proceedings of the 31st International Conference on Machine Learning (ICML) , 2014.
- [64] Martin Zinkevich, Markus Weimer, Alexander J Smola, et al. Parallelized Stochastic Gradient Descent [C]. In Advances in Neural Information Processing Systems (NIPS) , 2010.

- [65] R Gemulla, E Nijkamp, P J Haas, et al. Large- scale matrix factorization with distributed stochastic gradient descent[R]. In KDD , 2011.
- [66] Tianbao Yang. Trading Computation for Communication: Distributed Stochastic Dual Coordinate Ascent [C]. In Advances in Neural Information Processing Systems (NIPS) , 2013.
- [67] Martin Jaggi, Virginia Smith, Martin Takc, et al. Communication-Efficient Distributed Dual Coordinate Ascent[C]. In Advances in Neural Information Processing Systems (NIPS) , 2014.
- [68] Chenxin Ma, Virginia Smith, Martin Jaggi, et al. Adding vs. Averaging in Distributed Primal- Dual Optimization[C]. In Proceedings of the 32st International Conference on Machine Learning (ICML) , 2015.
- [69] A Nemirovski, A Juditsky, G Lan, et al. Robust stochastic approximation approach to stochastic programming[J]. SIAM Journal on Optimization (SIOPT) , 2009, 19(4) : 1574-1609.
- [70] T Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]. In Proceedings of the 21st International Conference on Machine Learning (ICML) , 2004.
- [71] M Zinkevich. Online convex programming and generalized infinitesimal gradient ascent[C]. In Proceedings of the 20th International Conference on Machine Learning (ICML) , 2003.
- [72] E Hazan, A Agarwal, S Kale. Logarithmic regret algorithms for online convex optimization[J]. Machine Learning , 2007, 69(2-3) : 169-192.
- [73] G Lugosi. Concentration- of- measure inequalities [R]. Department of Economics, Pompeu Fabra University, 2009.
- [74] A Nemirovski, D B Yudin. Problem complexity and method efficiency in optimization[M]. John Wiley & Sons Ltd, 1983.
- [75] A Agarwal, P L Bartlett, P Ravikumar, et al. Information theoretic lower bounds on the oracle complexity of stochastic convex optimization[J]. IEEE Transactions on Information Theory (TIT) , 2012, 58(5) : 3235-3249.
- [76] A Iouditski, Y Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions[J / OL] . 2010.
- [77] E Hazan, S Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization [C]. In Proceedings of the 24th Annual Conference on Learning Theory (COLT) , 2011.
- [78] A Rakhlin, O Shamir, K Sridharan. Making gradient descent optimal for strongly convex stochastic optimization[C]. In Proceedings of the 29th International Conference on Machine Learning (ICML) , 2012.
- [79] O Shamir, T Zhang. Stochastic gradient descent for non- smooth optimization: convergence results and optimal averaging schemes[C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [80] Y Nesterov. Introductory lectures on convex optimization: a basic course [M]. Kluwer Academic Publishers, 2004.
- [81] Y Nesterov. Smooth minimization of non- smooth functions [J]. Mathematical Programming , 2005, 103 (1) : 127-152.
- [82] Y Nesterov. Gradient methods for minimizing composite functions[J]. Mathematical Programming , 2013, 140(1) : 125-161.
- [83] A Cotter, O Shamir, N Srebro, et al. Better mini-batch algorithms via accelerated gradient methods[C]. In Advances in Neural Information Processing Systems 24 (NIPS) , 2011.

- [84] L Zhang, T Yang, R Jin, et al. $O(\log T)$ projections for stochastic optimization of smooth and strongly convex functions[C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [85] E Hazan, S Kale. Projection-free online learning[C]. In Proceedings of the 29th International Conference on Machine Learning (ICML) , 2012.
- [86] M Jaggi. Revisiting Frank-Wolfe: projection-free sparse convex optimization[C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [87] H Ouyang, N He, L Q Tran, et al. Stochastic alternating direction method of multipliers [C]. In Proceedings of the 30th International Conference on Machine Learning (ICML) , 2013.
- [88] S-Y Zhao, W-J Li, Z-H Zhou. Scalable stochastic alternating direction method of multipliers[J/OL]. arXiv: 1502.03529 , 2015.
- [89] N L Roux, M Schmidt, F Bach. A stochastic gradient method with an exponential convergence rate for finite training sets[C]. In Advances in Neural Information Processing Systems 25 (NIPS) , 2012.
- [90] S Shalev-Shwartz, T Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization [J]. Journal of Machine Learning Research (JMLR) , 2013 , 14: 567-599.
- [91] S Shalev-Shwartz, T Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization[J/OL]. arXiv: 1309.2375 , 2013.
- [92] L Zhang, M Mahdavi, R Jin. Linear Convergence with Condition Number Independent Access of Full Gradients[C]. In Advances in Neural Information Processing Systems 26 (NIPS) , 2013.
- [93] M Mahdavi, L Zhang, R Jin. Mixed Optimization for Smooth Functions [C]. In Advances in Neural Information Processing Systems 26 (NIPS) , 2013.
- [94] R Johnson, T Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction[C]. In Advances in Neural Information Processing Systems 26 (NIPS) , 2013.
- [95] J Konecny, P Richtarik. Semi-Stochastic Gradient Descent Methods[J/OL]. arXiv: 1312.1666 , 2013.
- [96] Q Lin, Z Lu, L Xiao. An accelerated proximal coordinate gradient method[C]. In Advances in Neural Information Processing Systems 27 (NIPS) , 2014.
- [97] A Nitanda. Stochastic proximal gradient descent with acceleration techniques[C]. In Advances in Neural Information Processing Systems 27 (NIPS) , 2014.
- [98] Y Zhang, X Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization [C]. In Proceedings of the 32nd International Conference on Machine Learning (ICML) , 2015.
- [99] Y Zhen, D-Y Yeung. A probabilistic model for multimodal hash function learning[C]. In Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) , 2012.
- [100] Y Zhen, D-Y Yeung. Co-regularized hashing for multimodal data[C]. In Advances in Neural Information Processing Systems 25 (NIPS) , 2012.
- [101] Y Wei, Y Song, Y Zhen, et al. Scalable heterogeneous translated hashing[C]. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD) , 2014.
- [102] W Kong, W-J Li Isotropic hashing [C]. In Advances in Neural Information Processing Systems 25 (NIPS) , 2012.
- [103] Q-Y Jiang, W-J Li. Scalable Graph Hashing with feature transformation[C]. Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI) , 2015.
- [104] P Zhang, W Zhang, W-J Li, et al. Supervised hashing with latent factor models[C]. In Proceedings of

- the 37th ACM Conference on Research and Development in Information Retrieval (SIGIR) , 2014.
- [105] D Zhang , L-W Li . Large- scale supervised multimodal hashing with semantic correlation maximization [C]. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI) , 2014.
- [106] W Kong , W-J Li . Double-bit quantization for hashing[C]. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI) , 2012.
- [107] W Kong , W-J Li , M Guo . Manhattan hashing for large- scale image retrieval[C]. In Proceedings of the 35th ACM Conference on Research and Development in Information Retrieval (SIGIR) , 2012.
- [108] B Xu , J Bu , Y Lin , et al . Harmonious hashing[C]. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI) , 2013.
- [109] C Wu , J Zhu , D Cai , et al . Semi- supervised nonlinear hashing using bootstrap sequential projection learning[J]. IEEE Transactions on Knowledge and Data Engineering (TKDE) , 2013 , 25 (6) : 1380-1393.
- [110] Q Zhang , Y Wu , Z Ding , et al . Learning hash codes for efficient content reuse detection [C]. In Proceedings of the 35th ACM Conference on Research and Development in Information Retrieval (SIGIR) , 2012.
- [111] D Zhai , H Chang , Y Zhen , et al . Parametric local multimodal hashing for cross- view similarity search [C]. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI) , 2013.
- [112] M Kan , D Xu , S Shan , et al . Semi-supervised hashing via kernel hyperplane learning for scalable image search[J]. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) , 2014 , 24 (4) : 704-713.
- [113] Y-M Zhang , K Huang , G Geng , et al . Fast kNN graph construction with locality sensitive hashing[C]. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD) , 2013.
- [114] P Li , M Wang , J Cheng , et al . Spectral Hashing With Semantically Consistent Graph for Image Indexing [J]. IEEE Transactions on Multimedia (TMM) , 2013 , 15 (1) : 141-152.
- [115] H Yang , X Bai , J Zhou , et al . Adaptive Object Retrieval with Kernel Reconstructive Hashing[C]. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2014.
- [116] C Leng , J Wu , J Cheng , et al . Hashing for Distributed Data [C]. In Proceedings of The 32nd International Conference on Machine Learning (ICML) , 2015.
- [117] M Ou , P Cui , F Wang , et al . Comparing apples to oranges: A scalable solution with heterogeneous hashing[C]. In Proceedings of the 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) , 2013.
- [118] Jile Zhou , Guiguang Ding , Yuchen Guo . Latent semantic sparse hashing for cross-modal similarity search [R] . SIGIR 2014.
- [119] Guiguang Ding , Yuchen Guo , Jile Zhou . Collective Matrix Factorization Hashing for Multimodal Data [R] . CVPR 2014.
- [120] F Wu , Z Yu , Y Yang , et al . Sparse multi- modal hashing [J]. IEEE Transactions on Multimedia (TMM) , 2014 , 16 (2) : 427-439.
- [121] H Peng , C Deng , L An , et al . Learning to multimodal hash for robust video copy detection [R] . ICIP 2013.
- [122] D Wang , X Gao , X Wang , et al . Semantic Topic Multimodal Hashing for Cross-media Retrieval[C]. In

- Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) , 2015.
- [123] Jingkuan Song, Yi Yang, Xuelong Li, et al. Robust Hashing With Local Models for Approximate Similarity Search[J]. IEEE T. Cybernetics, 2014, 44(7) : 1225-1236.
- [124] Lin Chen, Dong Xu, Ivor Wai-Hung Tsang, et al. Spectral Embedded Hashing for Scalable Image Retrieval[J]. IEEE T. Cybernetics, 2014, 44(7) : 1180-1190.
- [125] L-K Huang, Q Yang, W-S Zheng. Online hashing[C]. In Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI) , 2013.
- [126] F Shen, C Shen, Q Shi, et al. Inductive hashing on manifolds[C]. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2013.
- [127] Y Lin, R Jin, D Cai, et al. Random projection with filtering for nearly duplicate search [C]. In Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI) , 2012.
- [128] L Liu, P Fieguth, G Kuang, et al. Sorted random projections for robust texture classification[C]. In Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV) , 2011.
- [129] M Yuan, K Deng, J Zeng, et al. OceanST: a distributed analytic system for large-scale spatiotemporal mobile broadband data[C]. Proceedings of the VLDB Endowment, 2014, 7(13) : 1561-1564.
- [130] R Chen, J Shi, Y Chen, et al. PowerLyra: differentiated graph computation and partitioning on skewed graphs[C]. In Proceedings of the 10th European Conference on Computer Systems (EuroSys) , 2015.
- [131] C Xie, L Yan, W-J Li, et al. Distributed power-law graph computing: theoretical and empirical analysis [C]. In Advances in Neural Information Processing Systems 27 (NIPS) , 2014.
- [132] M Xu, B Lakshminarayanan, Y W Teh, et al. Distributed bayesian posterior sampling via moment sharing[C]. In Advances in Neural Information Processing Systems 27 (NIPS) , 2014.
- [133] J Zhu, J Chen, W Hu. Big Learning with Bayesian Methods[J/OL]. arXiv: 1411.6370 , 2014.
- [134] Bojun Tu, Zhihua Zhang, Shusen Wang, et al. Making Fisher Discriminant Analysis Scalable[R]. The International Conference on Machine Learning (ICML) , 2014.
- [135] Z-Q Yu, X-J Shi, L Yan, et al. Distributed stochastic ADMM for matrix factorization[C]. In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM) , 2014.
- [136] L Yan, W-J Li, G-R Xue, et al. Coupled group lasso for web-scale CTR prediction in display advertising [C]. In Proceedings of the 31st International Conference on Machine Learning (ICML) , 2014.
- [137] C Hu, J T Kwok, W Pan. Accelerated gradient methods for stochastic optimization and online Learning [C]. In Advances in Neural Information Processing Systems 22 (NIPS) , 2009.
- [138] L W Zhong, J T Kwok. Fast stochastic alternating direction method of multipliers[C]. In Proceedings of the 31st International Conference on Machine Learning (ICML) , 2014.
- [139] M Lin, S Weng, C Zhang. On the sample complexity of random fourier features for online learning: how many random fourier features do we need[J]. ACM Transactions on Knowledge Discovery from Data (TKDD) , 2014, 8(3) : 1-19.
- [140] W Zhang, L Zhang, Y Hu, et al. Sparse learning for stochastic composite optimization [C]. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI) , 2014.
- [141] P Gong, J Ye. Linear Convergence of Variance-Reduced Stochastic Gradient without Strong Convexity[J/OL]. arXiv: 1406.1102 , 2014.
- [142] SZhang, A Choromanska, Y LeCun. Deep learning with elastic averaging SGD[J/OL]. arXiv: 1412.6651 , 2014.

- [143] R Ge, F Huang, C Jin, et al. Escaping from saddle points—online stochastic gradient for tensor decomposition [C]. In Proceedings of the 28th Conference on Learning Theory (COLT), 2015.
- [144] 李武军, 周志华. 大数据哈希学习: 现状与趋势 [J]. 科学通报, 2015, 60(5-6): 485-490.

作者简介

周志华 南京大学教授, 中国计算机学会人工智能与模式识别专委会主任, 中国人工智能学会机器学习专业委员会主任。ACM 杰出科学家, IEEE Fellow, 中国计算机学会会士。



李武军 南京大学计算机科学与技术系副教授, 博士生导师。2010 年毕业于香港科技大学计算机科学与工程系, 获工学博士学位。主要研究方向为机器学习与数据挖掘。在 Artificial Intelligence、IEEE Transactions TKDE、ICML、NIPS、SIGIR、IJCAI、AAAI 等国际知名期刊和会议上发表论文 30 余篇。



张利军 南京大学计算机科学与技术系副教授, 硕士生导师。2012 年毕业于浙江大学计算机科学与技术系, 获工学博士学位。主要研究方向为大规模机器学习及优化, 在国际学术会议和期刊上发表论文共 30 余篇, 包括 ICML、NIPS、COLT、AAAI、ACM MM、AISTATS、TPAMI、TIT、TIP、TKDE。



新型数据管理系统研究进展与趋势

数据库专委会

摘要

随着大数据时代的到来，以关系型数据库为代表的传统数据管理技术受到了严峻的挑战。针对大数据的应用特点，产生了一批具有代表性的新型数据管理系统。本文介绍了当前新型数据管理系统的主要架构，系统总结了各种典型的系统，对国内外研究现状进行了对比，并对未来的发展趋势进行了预测。

关键词：数据管理系统，图数据，流处理，众包，在线分析，NewSQL

Abstract

With the arrival of the era of big data, the traditional data management technology represented by relational databases has been a severe challenge. A number of representative new data management systems are generated for supporting application features of big data. This paper introduces the main architecture of the new data management systems, summarizes various typical systems, compares the research status at home and abroad, and forecasts the future development trend.

Keywords: data management system, graph data, stream data, crowdsourcing, online analysis, NewSQL

1 概述

1.1 引言

随着人类进入 21 世纪，尤其是互联网和移动技术的发展，使得人与人之间的联系日益密切，社会结构日趋复杂，生产力水平得到极大提升，人类创造性活力得到充分释放，与之相应的数据规模和处理系统发生了巨大改变，从而催生出当下众人热议的大数据局面。

从历史观的角度看，数据 (D) 和社会 (S) 形成一定的对应关系，即： $D_1 \sim f(S_{\text{Sumerians}})$, ..., $D_{\text{big}} \sim f(S_{\text{present}})$, ..., $D_n \sim f(S_{\text{future}})$ 。从量的关系上， D_1 , ..., D_{big} , ..., D_n 可能存在大小关系，还可形成包含关系，但它们只是与当时的社会发展状况相对应： D_{big} 不可能反映代表未来的 D_n ，因为我们不知道未来会有什么新的社会结构（诸如当下社交网络一类的事物）出现，也不知道会有什么新的生产活动（诸如电商一

类的事物)产生;同样 D_1 也不需要具有 D_{big} 的规模,当时人们并没有如此频繁的联系。近期,美国加州大学伯克利分校Michael I. Jordan教授提出的“大数据的冬天即将到来”,如果我们能从历史的角度认识 D_{big} 的地位,没有把 D_{big} 当 D_n ,就不存在“冬天”与“春天”的问题。这是历史客观发展的事实。

基于以上分析,当下大数据的产生主要源于人类社会生活网络结构的复杂化、生产活动的数字化、科学的研究信息化相关,其意义和价值在于如何帮助人们解释复杂的社会行为和结构,以及提高人们生产制造的能力,进而丰富人们发现自然规律的手段。本质上,大数据具有以下三方面的内涵,即:大数据的“深度”、大数据的“广度”,以及大数据的“密度”。所谓“深度”是指单一领域数据汇聚的规模,可以进一步理解为数据内容的“维度”。数据的“广度”则是指多领域数据汇聚的规模,侧重体现在数据的关联、交叉和融合等方面。大数据的“密度”是指时空维上数据汇聚的规模,即数据积累的“厚度”以及数据产生的“速度”等。

面对不断涌现的大数据应用,数据库乃至数据管理技术面临新的挑战。传统的数据库技术侧重考虑数据的“深度”问题,主要解决数据的组织、存储、查询和简单分析等问题。其后,数据管理技术在一定程度上考虑了数据的“广度”和“密度”问题,主要解决数据的集成、流处理、图结构等问题。这里提出的大数据管理是要综合考虑数据的“广度”、“深度”、“密度”等问题,主要解决数据的获取、抽取、集成、复杂分析、解释等技术难点。因此,与传统数据管理技术相比,新型数据管理技术难度更高,处理数据的“战线”更长。

新型数据管理系统正是基于上述新的挑战,着重解决数据在“广度”、“深度”或者“密度”等方面的问题。本报告涉及的新型数据管理系统可以按照解决问题的侧重点做如下的简单分类:

表1 新型数据管理系统分类

系统类型	代表性系统	主要解决的问题
图数据管理系统	Prege、Giraph、PowerGraph、GraphChi、XStream、Giraph +	数据的“广度”和“深度”
流数据管理系统	S4、Storm、Puma、Samza	数据的“密度”
众包数据管理系统	Mturk、CrowdFlower、samasource、CloudCrowd	数据的“广度”
在线分析管理系统	Dremel、Drill、Impala、BlinkDB	数据的“广度”
商业数据管理系统	Oracle、MySQL、SQL Server、MongoDB、HBase、Cassandra、Redis、Spanner、OceanBase、RDS、Cloud SQL、Azure、RDS、BigTable、DynamoDB、SimpleDB	数据的“广度”

1.2 新型数据管理系统的处理架构

无论是工业界还是学术界,都已经广泛使用高级集群编程模型来处理日益增长的数据,如MapReduce。这些系统将分布式编程简化为自动提供位置感知(locality-aware)调度、容错以及负载均衡,使得大量用户能够在商用集群上分析庞大的数据集。

大多数现有的集群计算系统都是基于非循环的数据流模型（acyclic data flow model）。从稳定的物理存储（如分布式文件系统）中加载记录，一组确定性操作构成一个有向无环图（Directed Acyclic Graph, DAG），记录被传入这个 DAG，然后写回稳定存储。通过这个 DAG 数据流图，运行时自动完成调度工作及故障恢复。

尽管非循环数据流是一种很强大的抽象方法，但有些应用仍然无法使用这种方式描述，包括：1) 机器学习和图应用中常用的迭代算法（每一步对数据执行相似的函数）；2) 交互式数据挖掘工具（用户反复查询一个数据子集）。基于数据流的架构也不明确支持这种处理，所以需要将数据输出到磁盘然后在每次查询时重新加载，从而带来较大的开销。

当前新型数据管理系统的发展趋势主要有两个方向，如图 1 所示，一种是以 Hadoop 和 MapReduce 为代表的批处理（Batch Processing）系统，另一种是为各种特定应用开发的流处理（Stream Processing）系统，批处理是先存储后处理（Store-then-process），而流处理则是直接处理（Straight-through processing）。

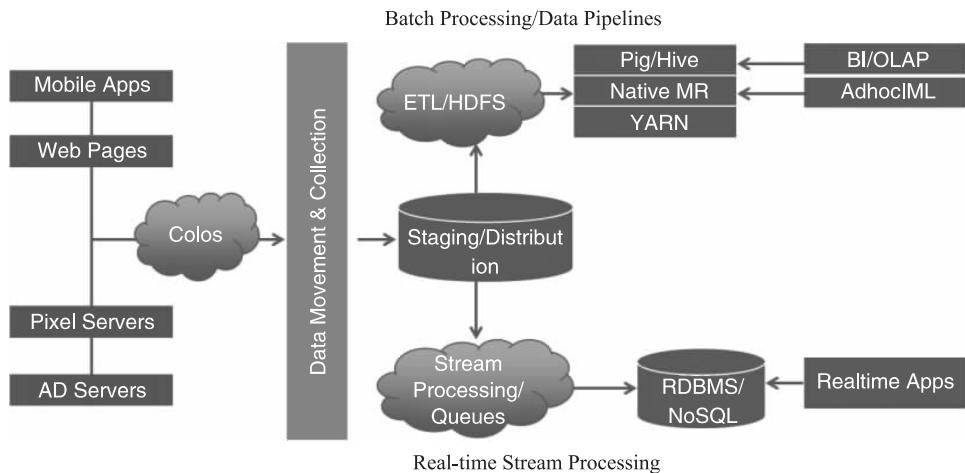


图 1 新型数据管理系统框架

随着大数据时代的到来，单一的计算模式已难以满足整个应用的需求，因此需要考虑不同计算模式的混合使用。Spark 作为混合式计算模式的典型代表应运而生，近年来逐渐引起了学术界和工业界的广泛重视。本章首先简单介绍大数据两种代表性系统处理架构：批处理模式和流处理模式，然后介绍以 Spark 为代表的混合处理模式。

1.2.1 批处理模式

Google 公司在 2004 年提出的 MapReduce 编程模型是最具代表性的批处理模式。一个完整的 MapReduce 过程如图 2 所示。

MapReduce 模型首先将用户的原始数据源进行分块，然后分别交给不同的 Map 任务区处理。Map 任务从输入中解析出键值对集合，然后对这些集合执行用户自行定义的 Map 函数得到中间结果，并将该结果写入本地硬盘。Reduce 任务从硬盘上读取数据之后，

会根据键值进行排序，将具有相同键值的数据组织在一起。最后用户自定义的 Reduce 函数会作用于这些排好序的结果并输出最终结果。

从 MapReduce 的处理过程我们可以看出，MapReduce 的核心设计思想在于：1) 将问题分而治之；2) 把计算推到数据而不是把数据推到计算，有效避免数据传输过程中产生的大量通信开销。MapReduce 模型很简单，且现实中很多问题都可用 MapReduce 模型来表示。因此该模型公开后，立刻受到极大的关注，并在生物信息学、文本挖掘等领域得到广泛应用。

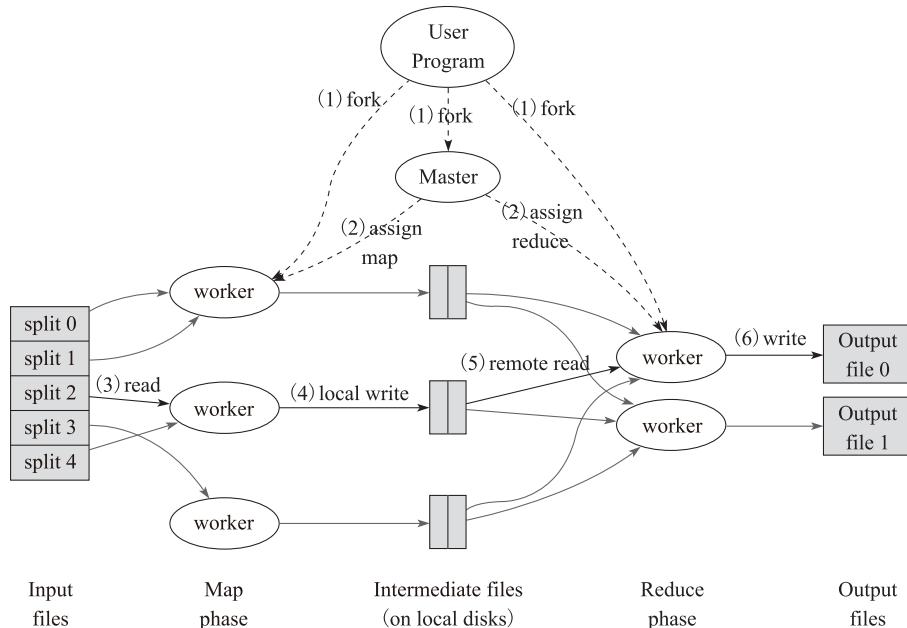


图 2 MapReduce 执行流程图

无论是流处理还是批处理，都是大数据处理的可行思路。大数据的应用类型很多，在实际的大数据处理中，常常并不是简单地只使用其中的某一种，而是将二者结合起来。互联网是大数据最重要的来源之一，很多互联网公司根据处理时间的要求将自己的业务划分为在线（Online）、近线（Nearline）和离线（Offline）。这种划分方式是按处理所耗时间来划分的。其中在线的处理时间一般在秒级，甚至是毫秒级，因此通常采用上面所说的流处理。离线的处理时间可以以天为基本单位，基本采用批处理方式，用这种方式可以最大限度地利用系统 I/O。近线的处理时间一般在分钟级或者是小时级，对其处理模型并没有特别的要求，可以根据需求灵活选择，但在实际中多采用批处理模式。

1.2.2 流处理模式

流处理的基本理念是数据的价值会随着时间的流逝而不断减少。因此尽可能快的对最新的数据做出分析并给出结果是所有流数据处理模式的共同目标。需要采用流数据处理的大数据应用场景主要有网页点击数的实时统计、传感器网络、金融中的高频交易等。

流处理的处理模式将数据视为流，源源不断的数据组成了数据流。当新的数据到来

时就立刻处理并返回所需的结果。图 3 是流处理中基本的数据流模型：

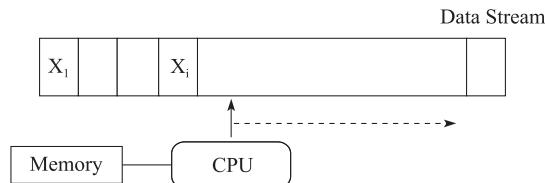


图 3 基本的数据流模型

数据的实时处理是一个很有挑战性的工作，数据流本身具有持续到达、速度快且规模巨大等特点，因此通常不会对所有的数据进行永久化存储，而且数据环境处在不断的变化之中，系统很难准确掌握整个数据的全貌。由于响应时间的要求，流处理的过程基本在内存中完成，其处理方式更多地依赖于在内存中设计巧妙的概要数据结构（Synopsis data structure），内存容量是限制流处理模型的一个主要瓶颈。以 PCM（相变存储器）为代表的 SCM（StorageClass Memory，储存级内存）设备的出现或许可以使流处理模型摆脱内存对其的制约。

1.2.3 混合处理模式

Spark 是混合处理模式的系统代表，是一种与 Hadoop 相似的开源集群计算环境，现在是 Apache 孵化的顶级项目，可用来构建大型的、低延迟的数据分析应用程序。Spark 启用了内存分布数据集，除了能够提供交互式查询外，它还可以优化迭代工作负载。

混合计算模式可体现在两个层面：一是传统并行计算所关注的体系结构与底层并行程序设计语言层面计算模式的混合，例如，在体系结构层，可根据大数据应用问题的需要搭建混合式的系统架构，如 MapReduce 集群 + GPU 的混合，或者 MapReduce 集群 + 众核协处理器的 OpenMP/MPI 的混合模型。二是大数据处理高层计算模式的混合。例如，一个大数据应用可能需要提供流式计算模式以便接受和处理大量流式数据；可能还需要提供基于 SQL 或 NoSQL 的数据查询分析能力以便进行日常的数据查询分析；并且可能需要提供线下批处理和迭代计算以完成机器学习的深度数据挖掘分析；一些大数据计算任务可能还涉及复杂图计算或间接转化为图计算问题等。

因此，很多大数据处理问题将需要混合使用多种计算模式。此外，为了提高计算性能，各种计算模式还可以与内存计算模式混合，实现高实时性的大数据查询和计算分析。混合计算模式将成为满足多样性大数据处理和应用需求的有效手段。混合计算模式集大成者当属 UC Berkeley AMPLab 实验室推出的 Spark 系统，如图 4 所示，其涵盖了几乎所有典型的大数据计算模式，包括迭代计算、批处理计算、内存计算、流式计算（StreamingSpark）、数据查询分析

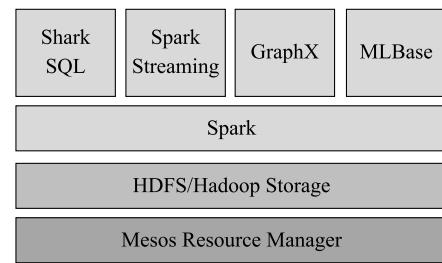


图 4 Spark 的体系结构

计算（Shark）以及图计算（Bagel、GraphX）。

1.3 总结

随着大数据时代的来临，新型数据管理系统需要关注数据的“深度”、“广度”和“密度”等维度，处理的模式主要包括批处理、流处理和混合处理三种模式。按照新型数据管理系统关注维度与处理方式的不同，本报告将数据管理系统分为图数据管理系统、流数据管理系统、众包数据管理系统、在线数据分析与管理系统和商业数据管理系统五种系统。以下章节将对各种系统的特点、国内外研究现状，以及今后的研究展望等方面展开对比分析。

2 图数据管理系统

2.1 引言

随着社交网络分析、Web语义分析、生物信息网络分析以及交通导航等新兴应用的快速增长，不同领域出现了规模庞大、内部结构复杂、查询需求多样的大图数据。2015年最新公布的统计数据^②显示，全球最大的社交网络Facebook拥有14.15亿月活跃用户，用户之间的好友关系超过2000亿条边，日均记录分享量也达到了数十亿级别，65%以上用户为日活跃用户；国内领先的QQ社交网站注册用户数也超过了8.29亿。

不同领域大图数据应用过程中存在通用的数据操作。例如，在社交网络中，发现影响力最大的用户，能够更加有效地推送广告；在交通网络中，发现两点间满足不同标准的路径，有助于客户行程规划；在金融交易网络中，发现特定的模式，有助于发现异常的交易行为；在社交网络中，发现关联关系密切、特征相近的用户群体，有助于商品的销售。这些大图数据操作不仅仅涉及数据节点自身的信息，而且涉及数据节点之间的结构关系。我们可以将大图数据操作抽象为获取特定节点，获取特定路径、获取特定子图等，通过实现这些通用操作，构建大图数据管理系统。

图数据管理在数据库领域是一个经典的研究问题，其研究历史呈现一个波浪式前进的过程。早在20世纪70年代，由于图数据模型表达能力强，数据管理领域的研究人员就提出图模型对客观世界的数据进行建模，并设计了相关的图数据管理原型系统。Charles W. Bachman还由于其在图数据模型方面的贡献于1973年获得图灵奖。之后，由于图数据查询在表达和执行方面的复杂度都很高，图数据管理系统在应用方面存在挑战，研究趋缓。在这一阶段，关系数据库由于其操作接口简单，查询优化技术实现突破，逐

^② <http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>

渐成为数据管理中的主流。2000 年之后，随着社交网络等真实大图数据的迅猛增长和其上应用需求的推动，图数据的相关研究工作重新成为热点。例如，在 VLDB2014 国际会议中 (<http://www.vldb.org/2014/program/Menu.html>)，出现了 4 个图数据管理的专题讨论。国际 IT 巨头，包括 Google、Facebook、微软等正在进行分布式大图数据管理系统的研发，支持包括海量 Web 网页重要性排序、社区发现等操作。

图数据管理也得到计算机方向其他研究领域的关注。例如，在 2014 年系统软件专委会所撰写的《面向新型计算模型的系统软件研究新进展与趋势》文中，上海交通大学陈海波团队也将《图并行计算系统的研究进展与趋势》作为重要内容进行论述。本报告将侧重从数据管理的角度对大图数据管理的相关技术展开论述。部分关键技术或系统如在《图并行计算系统的研究进展与趋势》论述过，本报告将不再赘述。

2.2 大图数据管理面临的挑战

大图数据是大数据的重要形态，具有大数据的 4V 特性：首先大图数据规模庞大，不仅仅包含大量图数据节点及其节点属性，而且包含图数据节点之间复杂的关联关系；其次，大图数据表达灵活，不同类型节点结构异构，每个节点和边上属性模式信息各不相同，很难抽象出相对固定的模式。例如，在客户购物关系图中，每个客户或者商品的属性不同，客户（或者商品）内部存在关联关系，客户与商品之间也存在关联关系。再次，大图数据处于动态变化中。随着时间不断的变化和演进，大图的内容信息发生变化，大图的结构信息也发生变化，如节点/边的新增和删除等。最后，大图数据中也蕴含着丰富的价值，通过分析大图数据中的结构、关系、时序等信息，为不同领域的大图数据应用决策提供支持。

大图数据还具有复杂性 C 的特性。由于大图数据表达方式灵活，任意数据节点之间都可能存在关联关系，假定节点数量为 n ，单一类型关联关系最多为 n^2 ，多类型关系场景下，关联关系就更加丰富。整个大图看作一个复杂网络。大图数据操作往往涉及全局数据，计算代价高。

大图数据的 4V + C 的特点使得大图数据的管理面临巨大挑战。第一，大图数据环境中图操作代价高。和传统的关系数据或者 XML 树状数据相比，图数据缺乏结构约束，操作复杂。以图数据中的两点最短路发现为例，经典 Dijkstra 算法的复杂性为 $O(m + n \log n)$ ，其中 n 是图中的节点数， m 是边数。随着图规模的增长，上述操作的代价急剧上升。此外，这一算法复杂性基于单机内存环境。考虑到单机内存无法满足大图数据处理需求，大图数据操作还需要考虑磁盘访问代价和分布式环境中网络访问代价。

第二，大图操作需求灵活，提供简洁、高效、表达能力强的查询语言面临困难。随着大图数据应用的深化，大图数据之上的操作需求日益复杂。大图数据操作不仅仅局限于类似检索某个点的所有关联边这样简单的查询，而且还包括更加复杂的、侧重于全局的查询和分析，如图中最短路发现查询、最小支撑树构建查询、子图模式匹配等操作。考虑到大图数据操作的灵活性，很难抽取公共部分进行优化或者简化表达。此外，图数

据操作通常需要循环迭代，这对方便、易用的图查询语言的设计有很大的影响。

第三，大图数据复杂性的特点使得很难直接应用并行计算模型解决大图数据操作。图数据规模庞大使得并行分布策略不可避免。同时，由于图数据内部关联灵活、复杂，图数据操作很可能需要全局的图数据信息。即使实现图数据的分块和并行操作，不同分块之间依然存在大量的访问，无法简单直接实施并行操作。

2.3 大图数据管理中的主要研究问题

面对大图数据管理的挑战，很多研究机构从不同的角度展开研究工作，包括大图数据管理的体系结构、大图数据的组织、大图数据管理的查询设计、大图数据查询执行和优化等问题。

2.3.1 大图数据管理的体系结构

在大图数据管理中，不同的底层体系结构、底层系统软件、硬件都会对大图数据管理产生影响。

分布式大图数据管理是大图数据管理的主流结构。由于大图数据的规模庞大、节点信息丰富、计算代价高，利用多机的计算资源和存储资源管理大图是大图数据管理的趋势。在分布式大图数据管理中，目前应用比较广泛的是基于 MapReduce 和基于 BSP 计算模型的大图数据管理。基于 MapReduce 框架实现大图数据管理是可行的技术路线。MapReduce 框架是目前通用大数据的并行处理框架，大图数据的查询可以通过 MapReduce 框架实现，具有良好的可扩展性和大数据分析的高吞吐量。但是，由于 MapReduce 框架处理大图数据导致过多的磁盘读写操作，基于 MapReduce 框架的大图数据管理效率不高。

基于 BSP 计算模型的大图数据管理方案具有较好的可扩展性和较高的效率。为了克服 MapReduce 管理大图导致的各类问题，Google 借鉴 BSP 计算模型的思路，设计了 Pregel 框架。Pregel 框架将图数据存储在分布式环境中不同计算节点的内存中，支持以点为中心的计算模式，通过内存随机访问直接修改节点的状态，避免磁盘读写的 I/O 代价。Google 的 Pregel 系统已经用于 Web 级别的大图节点重要性排序等操作，展示出较强的可扩展性。Pregel 在 Apache 社区的开源版本 Giraph 在 Facebook 也得到重要应用，在社区发现、子图聚类等操作上取得良好效果。

单机大图数据处理是一种重要的技术路线。随着 CPU 和内存技术的发展，使得单机有资源支持海量大图的处理。单机环境中无需考虑分布式大图数据管理中引入的网络代价，其算法设计实现简单，调试安装代价相对低廉。为了提高大图数据处理的可扩展性，而单机内存有限，单机计算环境中使用外存不可避免，单机大图数据处理框架的一个关键问题就是需要降低外存随机扫描对大图数据操作的影响。目前单机大图数据计算框架，如 GraphChi、XStream 等，处理常用的测试图数据（如 Twitter 数据），能够取得和分布式大图数据计算框架可比较的性能。

基于关系数据库管理大图数据是另一种可选的技术方案。关系数据库在联机事务处

理领域取得巨大成功，成为企业信息基础设施中不可或缺的组成部分，有巨大的市场份额。图数据管理和关系数据管理也有很多相通之处，如都需要数据存储、数据缓存、索引、查询优化、并发控制等。基于关系数据库能够快速实现超过内存的大图数据的管理，如索引、更新、并发操作等。关系数据库管理图数据面临的问题是关系数据库擅长一次处理一个集合，而图数据操作考虑优化策略，往往采用一次一个节点的操作方式，直接利用关系数据库管理大图数据效率不高。

充分利用新硬件的特性是提高图数据管理系统性能的有效手段。CPU 中 Cache 是影响计算性能的重要因素。在图数据组织和操作过程中考虑 Cache 的特性，提升 CPU 的 Cache 命中率，能够在全内存环境中进一步提高系统性能。图形处理器 GPU 带有大量的并行计算模块，通过设计合适的图数据分块和并行方法，结合 CPU 和 GPU 提高大图数据管理的效率。相对于带有机械设备的硬盘，采用电子寻址的闪存更加适合支持带有大量随机访问的图数据操作。在基于闪存的图数据管理中，需要考虑闪存写入代价昂贵的约束。

2.3.2 查询语言设计

图数据表示方式灵活，查询分析形式多样。图数据查询不仅包括对特定子图数据的获取，而且包括图数据的转换。由于查询方式多样，目前没有一种广为接受、易于表达、便于优化支持高效执行的查询语言。我们分析图数据管理中可能的查询语言方式。

提供图数据基本操作 API 是目前大图数据管理系统的一种查询方式。目前的图数据库管理系统，如 Neo4j，提供了基本的图操作 API，如图数据存储、遍历、最短路查询等操作。用户在其应用程序中仅仅需要调用这些 API，系统负责这些调用的执行和优化。然而由于图的操作方式灵活，利用有限的 API 操作很难支持不同领域的图操作需求。例如，Neo4j 无法直接实现 PageRank 的迭代计算。

根据现有图数据管理框架的接口要求，编写图数据不同查询脚本是目前图数据查询和分析的主流方式。系统不提供相对高层的图操作 API，而是提供相对底层的编程接口，如 MapReduce 框架接口、以点（边）为中心的计算接口等。用户根据查询流程，编写符合接口的查询脚本。这种方式足够灵活，能够支持复杂的图查询操作。但是，这种方式需要用户掌握脚本语言，负责脚本的编写、调试、优化等，用户应用开发负担重。

提供描述性图查询是介于图操作 API 和图框架接口的一种方式。类似于关系代数，图代数将不同类型的图操作公共部分抽取出来，形成图数据操作体系。用户通过相对高层的描述性查询表达图操作，系统负责将这些查询转换到底层框架中。这种方式能够屏蔽不同的底层平台的实现细节，减少用户应用系统实现的代价。然而，由于图操作本身的复杂性，即使采用描述性查询，描述性查询自身也很复杂，例如，描述性图查询中，需要有递归或者迭代机制，还需要选择、赋值等操作。

图的关键字查询是图数据查询的一种有前景的查询方式。图数据具有灵活的结构特性，用户通过结构信息表达查询繁琐。而现实中图数据是带有其他属性信息的。用户可以采用关键字描述图查询目标，图查询反馈包含这些关键字的特定子图，图的结构信息用于查询结果的排序。采用这种方式无需用户了解复杂的图结构，极大减轻用户图查询

的代价，但是不支持图数据的抽取和转换。

面向特定领域的图查询也是一种有效的图数据查询方式。特定领域的需求实际上是图数据查询的一种约束和限制。比较典型的领域相关查询语言是知识图谱领域中的 SPARQL。SPARQL 采用类 SQL 查询的形式表达查询模式，查询结果为与查询模式匹配的数据子图。SPARQL 采用类 SQL 的形式，能够降低用户的学习成本，提高用户的接受度。

2.3.3 大图数据组织

大图数据管理需要根据所在的计算环境，合理组织数据，提高数据处理的效率。在分布式环境中，大图数据需要划分到不同的计算节点中进行保存和处理。图数据划分需要考虑不同图划分之间的消息通信、计算节点的负载平衡等因素。最原始的图划分方法采用随机节点分块方法，根据 Hash 函数将图节点划分到不同的分区块中，或者采用节点 ID 区间的方式进行划分。由于不同子图之间的边意味着消息通信量，可以采用目前综合性能优异的 Metis 图划分方法，获得降低子图间通信量的图划分方法。由于同一数据节点在计算节点中出现并且仅出现一次，这种划分策略又称为点划分策略。

针对度数分布符合幂律分布的图数据，可以采用边划分的策略。即选择度数较大的数据节点进行切分，建立这一个数据节点在不同的计算节点中的副本，分散存储相关联的数据边，利用副本之间少量的消息量，降低点划分形成的子图间消息数量，减少采用点划分导致的计算负载不均衡的问题，提高整体运行效率。

现有系统还采用了以边为中心的计算模式。通常来讲，图中边的数量远大于点的数量，而边的信息在图数据操作中往往是无需改变的，点上附加的数据值是需要频繁改变的。以边为中心的思想将节点信息和边信息进行分离，将边的信息置于外存，节点信息尽量置于内存，通过顺序扫描边，改变内存中节点的数值。由于处理脚本主要是改变节点的数据值，这种方式本质上也是一种以节点为中心的计算。

在单机外存环境中，图数据划分内部还需要进一步组织，减少外存随机扫描对查询效率的影响。例如，按照边某一端点的 ID 实现图数据的顺序划分，而在同一个划分内部，按照另一端点的 ID 进行排序。这种组织方式能够在某一个图划分加载到内存时，通过顺序扫描特定的划分片段可以获取这一图划分相关边信息，减少外存随机访问的影响。

2.3.4 大图数据查询分析的执行和优化

目前分布式大图数据操作执行方式分为同步执行方式和异步执行方式。在同步执行方式中，同一个计算任务由多个超步组成。在每个超步中，节点接收到其他节点发送的消息，根据用户脚本处理消息，向其他节点发送消息。每个超步间通过栅栏保持同步，超步结束需要等待所有的参与计算节点结束。同步执行方式的优势为编程易于调试，系统稳定性高，可扩展性高。

与同步方式对应的执行方式为异步执行方式。其特点是执行过程中没有全局统一的同步概念。主要思路为数据节点在任何时刻，都可以进行计算。异步的方式优势在于能够快速收敛，执行效率较高。但是，异步执行方式无法保证执行过程中的确定性。即同

一任务多次执行，最终结果可能都不一样。这一特点在某些数据操作中影响计算有效性，并使得算法在分布式环境中难以调试。

减少分布式环境中通信量是大图数据操作的一种优化策略。传统以节点为中心的计算模式中，不同数据节点之间通过消息或者共享内存进行通信。如果能够感知到同一计算节点的其他数据节点，则可以通过效率较高的局部内存访问替代昂贵的网络访问。良好的图划分方法是实现这一优化的基础。

负载平衡是分布式查询执行过程中需要考虑的问题。分布式环境中同步执行查询的代价取决于最慢的计算节点。不同类型的图数据查询产生不同的负载分布。根据计算节点的实时负载，迁移数据节点，实现负载的动态平衡，能够查询的运行时间。在数据节点的动态迁移过程中，需要考虑数据节点在分布式环境中的重定位问题。

查询脚本层面的优化是大图查询优化的有效手段。图数据操作通常涉及循环迭代，而在分布式系统中每次循环都有额外代价，如 BSP 模型中超步的同步代价、MapReduce 框架中图数据的扫描和传输等。此外，某些循环在分布式环境中活跃节点过少，循环带来的相对收益差。减少分布式环境中的循环次数是图数据查询重要的优化策略。

2.4 国外研究现状

面对大图数据带来的技术挑战，国外高校、研究机构展开大图数据的研究工作，设计并实现了大图数据管理系统。我们以典型系统为代表，分析国外大图数据管理相关的研究进展。

2.4.1 大图数据分布式管理框架

Pregel: Pregel 是 Google 在 2010 年发表针对大图数据的分布式计算框架，是目前大图数据管理领域最有影响力的研究成果之一。针对 MapReduce 框架管理大图缺乏迭代支持、随机数据修改代价昂贵等问题，Pregel 框架做了创新和优化。Pregel 将不同数据节点进行分割，分别加载到不同的计算节点中。图数据查询整体遵循 BSP 计算模型，每个图查询由多个超步加以完成。每个超步中，图数据节点计算采用以节点为中心的编程模型。Pregel 框架在 Google 得到广泛应用，支持 Web 级别的大图数据节点重要性排序等基本操作。

Giraph: Giraph 是 Apache 所支持的，它是一个在 Hadoop 之上运行的 Pregel 的开源系统。Giraph 中实现了 BSP 的计算模型，支持用户以点为中心的脚本编程。相对于其他分布式图查询系统，Giraph 有一个显著特点，即 Giraph 本身是一个无 Reducer 的 MapReduce 任务，图数据查询处理易于和 Hadoop 环境进行集成。Giraph 在 Facebook 有重要的应用，支持社区发现、节点重要性排序等操作。

Trinity 系统：微软研究院设计了 Trinity 系统。该系统能够同时支持低延迟的在线查询和高吞吐量的离线分析。为了避免外存随机扫描的影响，Trinity 将图数据保存在内存云中，并且设计了复杂的内存机制来管理边长数据。Trinity 在上层引入了查询语言 TQL 表达图数据的查询。

GraphLab 和 *PoweGraph*: GraphLab 是一个支持异步执行方式、利用共享内存的大图数据管理框架。为了提高异步执行中的数据一致性, GraphLab 引入了数据节点访问的锁机制。针对自然图中高度数节点带来的负载偏斜, GraphLab 的后续版本 PowerGraph 引入了边划分的策略, 建立高度数节点的副本, 划分关联边到不同的计算节点, 使得每个计算节点上拥有均衡数量的边, 从而提高查询执行效率。

2.4.2 大图数据单机管理框架

GraphChi 和 *Xstream*: GraphChi 是单机图数据查询框架。单机环境处理大图不可避免需要外存的支持, 而大量随机访问外存数据会产生严重的性能问题。GraphChi 将图数据按照节点 ID 区间进行划分, 保存入边信息, 每个分区的出边信息在其他分区中顺序存储, 这样利用外存的顺序访问替代随机方法, 提升单机外存环境中以点为中心的计算框架的性能。XStream 采用以边为中心的存储方式, 不需要对边表进行排序, 也不需要索引, 同时通过顺序存取提高缓存的命中率。

2.4.3 基于关系数据库的大图管理框架

Vertexica: Vertexica 是 MIT 开发的基于关系数据库 (列式存储) 的图数据管理框架, 其特点是基于数据库系统存储图数据, 其上提供了以点为中心的计算模式, 支持用户通过编写脚本管理图数据; 还提供了一种查询语言 GRAPHiQL, 支持用户表达关系操作和图数据操作。

2.4.4 大图数据高层查询语言

Socialite: Socialite 是斯坦福大学开发的一种基于 Datalog 的大规模图数据查询系统。Socialite 扩展 Datalog 支持嵌套底层数据, 支持递归实现聚集操作, 允许用户自定义执行次序等。Socialite 可以方便表达可达性查询、最短路查询、PageRank、三角形计算等图查询。用户提交的 Socialite 查询能够自动转换到底层执行计划, 并实现优化。

表 2 部分大图数据管理框架特性

	代码	体系	同步/异步	数据访问	编程模型
Pregel	非开源	分布式	同步	消息	点中心
Giraph	开源	分布式	同步	消息	点中心
PowerGraph	开源	分布式	同步/异步	共享内存	点中心
GraphChi	开源	单机	同步	共享内存	点中心
XStream	开源	单机	同步	消息	边中心
Giraph +	非开源	分布式	同步/异步	消息/共享内存	子图中心

2.5 国内研究现状

下面我们从大图数据管理框架、大图数据操作算法、大图数据应用三个层面讨论国

内在大图数据管理方面的进展。

2.5.1 大图数据管理框架

BC-BSP: BC-BSP 是东北大学和中国移动合作开发的，基于 BSP 模型的大图数据分布式并行迭代计算的平台。BC-BSP 系统支持基于虚拟桶的均衡哈希划分、基于边聚簇特性的图划分动态调整策略，提供了以分区（子图）为中心和以（顶）点为中心的两种计算模式和编程接口，提高了数据处理的灵活性和适应性。与目前的开源框架 Hama 以及 Giraph 相比，BC-BSP 具有良好的可扩展性。

MoCGraph: MoCGraph 是北京大学开发的基于消息流式处理的可扩展大图查询框架，在 BSP 环境中能够以数据流的方式处理大图计算中的消息数据，减少中间结果对框架内存的压力。同时，利用流式计算减少针对外存的随机访问操作。

PSgL: PSgL 是北京大学开发的子图并行列举框架，支持在图数据中枚举所有与模式图同构的子图。其利用分治思想，基于图遍历操作迭代枚举结果，设计了多种分发策略保证计算节点间的负载均衡，提出了三种独立机制来缩小中间结果规模，支持索引查询、图结构分析等重要应用。

VENUS: VENUS 是华为研究人员所研制的单机大图数据管理框架。VENUS 针对 GraphChi 框架存在的问题进行了优化，分别存储支持更新节点信息和只读边信息，通过顺序扫描减少边数据对内存中的占用，通过将节点数据尽可能放入内存，减少访问节点数据对随机扫描的影响。

GraphHP: GraphHP 是西北工业大学开发的图数据管理框架，结合图查询执行中的同步模式和异步模式。GraphHP 中数据节点分为边界节点和局部节点。在同步环节，边界节点之间进行消息通讯，而局部节点的计算通过异步操作完成。通过两阶段的超步模式，加快了系统收敛速度。

GraphMemory: GraphMemory 是东北大学开发的图数据存储和查询优化系统，结合集群环境下的异构计算资源（如众核、多处理器、多处理节点等）以及内存访问特性，提供适合大图数据的并行处理框架、分析任务调度策略，以及系统负载平衡算法；采用基于世系的数据容错技术和基于热备进程的任务恢复技术，来提升整个系统的可靠性和可用性。

2.5.2 大图数据查询和分析

GStore 和 *GAnswer*: GStore 是北京大学开发的利用子图匹配技术面向海量 RDF 数据的 SPARQL 检索引擎。gStore 将 RDF 数据存储为图数据，利用图模式匹配操作支持图 SPARQL 查询，引入结构编码减少模式匹配过程中的搜索空间。gAnswer 引入了 RDF 数据的自然语言查询方式，避免用户编写复杂的结构查询，利用子图匹配，同时获取自然语言查询答案和短语消歧。

Laft-Suite: Laft-Suite 是清华大学研发面向社交网络分析系统，深入解决了社区结构分析、重要节点及潜在重要节点发现、社交链接产生方向推断、社交网络演化过程推断以及社交网络演化预测等问题，提供了一系列有效的解决方案，不仅支持传统的社交网

络静态结构分析，而且能展现社交网络的生成过程和演化规律。

2.5.3 大图数据应用系统

学术空间 (*ScholarSpace*)：学术空间是人民大学开发的，以学者为中心的个人学术信息空间。支持各领域基于作者名的学术信息查询，自动生成包括发表文献列表、发表数量曲线图、合作作者列表/图形化展示及知名学者图片、新闻的个人学术成就总页面。还提供了面向各学科领域的文献发表数量机构排名统计功能和各期刊收录文献单位和作者的 Top 20 排名统计功能，方便用户更直观地了解本领域的学术研究现状。同时系统按月提供所收录刊物的文献导读辑录，方便学者浏览本领域文献。系统通过自主研发的 Web 数据抽取和数据集成方法^[147,148]构建了语义丰富的关联数据库，目前具有近 200 万论文实体，100 多万作者实体。系统借助 RDF 通过三元组形式对关联数据进行描述，结构与大图数据十分类似。学术空间 ScholarSpace 采用北京大学 gStore 对 RDF 数据进行存储，包括 1 000 万条三元组、60 万节点，其查询引擎通过编码等方式对图的节点和边进行组织，并通过子图匹配算法实现图上的高效查询，具有良好的性能和可扩展性，能够在集群上进行实现，取得了良好的应用效果。

学者网：学者网是华南师范大学开发的，面向学者的社交网络，目前用户及学术数据记录已超过亿条，形成了复杂的图数据模型。基本存储模式采用关系数据库与分布式存储混合方式；对需要频繁读取的数据，实现高效的缓存查询和更新技术；对于需要进行大规模处理的数据，采用离线计算的方式实现；在进行论文搜索、学者关系和学术社区挖掘等推荐系统，采用了 Hadoop 分布式框架对数据进行处理。

2.6 国内外研究进展比较

面对大图数据管理的迫切需求，从工业界到学术界都开展了图数据管理的研究工作，图数据管理成为国内外的研究热点。国内的研究人员逐渐从大图数据算法研究，扩展到框架实现以及系统应用。我们从大图数据管理的框架层面和算法层面比较国内外的研究进展。

在框架层面，国外研究机构所提出的 Pregel、PowerGraph、GraphChi 等框架在系统成熟度、系统应用、原创性方面，要优于国内高校和科研院所所提到的框架。Pregel 在 Google 已经成功支持 Web 级别大图数据操作，PowerGraph 提供了多种图数据分析挖掘算法，GraphChi 在单机上取得集群可比较的性能，上述框架全部开源。国内各个高校和研究机构也在框架层面做了大量的工作，但是更多侧重于现有框架在某些方面的改进，包括消息处理机制、任务调度机制、图数据分割机制、查询优化机制等，期待更多原创性、系统性、实用性的成果出现。

在图数据分析算法层面，国内研究结构针对最短路发现、社区发现、SimRank 计算、模式匹配、信息推荐等热点问题，提出了多种图数据分析和挖掘方法，在方法的可扩展性、方法的效率，方法的资源消耗等方面，达到和国外同行所提出的方法相当的水平。

图数据研究的最大推动力是应用的发展。国内社交网络平台数据（如微信），购物网络平台数据（如淘宝）等发展迅猛，产生了大量的图数据，也带来了现实的挑战。华为、腾讯、百度、阿里等国内巨头已经开始了大图数据管理关键技术的研究和应用。这些基于真实场景和需求的研究更有生命力，将提高国内大图数据管理的研究水平。

2.7 总结

随着应用的逐步深化，图数据管理的研究也将继续发展。我们将图数据管理的发展趋势归纳为以下几点。

兼顾在线查询和离线分析的大图数据管理系统：不同类型的计算任务具有不同的特点，在线应用要求实时性高、响应及时，而离线处理任务更关注任务吞吐量、并发能力以及资源的合理利用性。目前所提出的图数据管理的框架，包括 Pregel、GraphLab、GraphChi 等，更多侧重于图数据的分析，强调图数据分析的吞吐量，查询的实时响应不是目前框架的重点。兼顾查询效率和吞吐量的大图数据管理系统将结合两种处理方式的优点，屏蔽具体的计算模型，根据用户提交的查询和数据量自动选择合适的处理方式。

支持丰富属性的大图数据管理框架：目前的大图数据处理框架，更多侧重于大图数据结构的分析，而现实世界中的大图数据往往带有属性信息。大图数据实际上包含结构和内容两类数据。大图数据之上仅考虑结构特性的查询应用场景较窄。充分借鉴其他研究领域，如信息检索领域、数据挖掘领域的研究进展，实现结构和内容相结合的大图数据查询和分析，为应用提供更有效的大图数据管理服务。

大图数据管理中的事务：随着大图数据应用的不断深入，多个用户可能同时操作大图数据，同时大图数据是保存在不同的数据中心中，如何支持多个用户并发访问大图数据，保证数据一致性是大图数据后续需要解决的问题。某些最新的工作已经开始探讨分布式大图计算环境中错误恢复等操作，后续还有广阔研究空间。

3 流数据管理系统

3.1 引言

近年来各行业信息化程度明显加快，由此产生的数据量也呈爆发式增长。在金融应用，网络监视，社交网络等行业领域产生了一类到达速度快，数据规模大的数据。这类数据的特点可以总结为^[86]：

- 1) 数据实时到达，到达速度较快。
- 2) 数据到达次序独立。

- 3) 数据规模庞大，无法预知数据的大小。
- 4) 数据一经处理，除非进行存储，否则很难再次获取。

我们把具有以上特征的一类数据称作流数据。自 20 世纪 90 年代开始，就有关于流数据管理的研究，当时主要应用在报警环境中^[28]。而后在不同的应用领域，逐渐产生了信息过滤系统^[29]、实时数据库^[30]、主动数据库^[31]等技术。21 世纪初，出现了 Aurora^[32]、TelegraphCQ^[33]、Stream^[34]等典型的数据流管理系统。这些早期的流数据管理系统提供基本算子和持续查询语言（CQL），但存在以下问题：

- 1) 系统通常是针对特定领域和场景，在通用性上较为欠缺。
- 2) 提供的查询处理功能较为单一，无法进行较为自由的数据处理操作。
- 3) 采用集中式的系统架构，对系统的扩展性和处理能力有较大的限制。

随着数据量的不断增长，传统的集中式的系统架构在处理能力上开始显现出不足。研究人员利用数据到达次序独立的特性，把分布式计算相关技术引入流数据管理系统。在集中式的流数据管理系统基础上，提出了 Medusa^[35]、Flux^[36]、Borealis^[37]等分布式流数据管理相关技术。

2004 年 MapReduce^[38] 计算模型的提出，以及随后 Hadoop^[39] 平台的出现和发展，为大数据环境下的批处理任务提供了较为有效的解决方案。但这种依赖磁盘存储的计算模式并不能满足流数据处理对实时性的要求。大数据领域的流数据处理开始逐渐受到关注和重视。

为了解决流数据处理面临的问题，提升流数据管理系统的计算能力和响应速度，研究者们开始采用分布式技术重新设计流数据管理系统，提出了 S4^[39]、Storm^[40] 等分布式流数据处理系统。与早期集中式流数据管理系统比，分布式流数据处理系统拥有更好的伸缩性和扩展性，提供了更加灵活的数据处理自由，能够对大量快速到来的数据进行有效的实时的处理^[41]。但分布式的整体结构也带来了故障恢复，负载均衡等一系列新的研究课题。

与此同时，越来越多的研究者基于现有的流数据管理系统，在流数据的环境中利用新的技术手段对流数据的挖掘和数据分析等问题进行研究，并取得了一定的成果。

本节将主要对国内外近年来在流数据处理系统上的发展及研究进展进行综述，并探讨未来流数据相关的研究和发展方向。

3.2 国际研究现状

3.2.1 国外典型分布式流数据处理系统介绍

随着大数据的发展，对海量数据的管理和分析受到越来越多的关注。流数据作为大数据中的一种特殊类型的数据，重要程度也逐渐凸显。从 Yahoo！开源第一个分布式流数据处理系统 S4 开始，越来越多的分布式流数据处理系统先后被提出。截止到目前，出现了 S4、Storm、Spark Streaming^[41]、Samza^[42]、MillWheel^[43]、Heron^[44] 等各具特色的分布

式流数据处理系统。下文将对这几种典型系统作分别介绍[⊖]。

表 3 典型流处理系统对比

	S4	Storm	Spark Streaming	Samza	MillWheel
提出时间	2010	2011	2012	2013	2013
系统整体结构	去中心化	弱中心化	中心化	中心化（依赖 YARN）	中心化
数据模型	event	tuple	object	object	$\langle \text{key}, \text{val}, \text{ts} \rangle$
处理粒度	单记录	单记录/批次	微批次	单记录	单记录/窗口
排序/去重	不提供	不提供	不提供	不提供	提供
数据路由	key 值路由	内置多种/自定义	/	依赖 Kafka	key 值路由
负载均衡	静态	非自适应动态	/	非自适应动态	自适应动态
反馈机制	无	提供	/	基于 Kafka 顺序 ACK	提供
语义保障	最多一次	精确一次	至少一次	至少一次	精确一次
状态存储	本地内存	本地内存/远程数据库	本地内存	本地内存/备份节点	后备存储
中间结果存储	内存队列	内存队列	本地内存/可靠文件系统	Kafka	内存/备份仓库
故障恢复	无状态被动备用	上游备份	检查点/并行恢复	上游备份	检查点

1. S4

S4 是由 Yahoo! 公司 2010 年开源的通用分布式流数据处理系统。平台的设计遵循 Actor 模式^[45]。系统节点间通过 Zookeeper^[46] 进行协调，采用去中心化的系统整体结构。S4 采用 PE 作为计算单元，把数据抽象为事件（event），事件通过 key 值进行划分和路由，每个 PE 只能处理同一 key 值的事件。去中心化的设计带来了较好的伸缩性与扩展性，但也因此存在一些局限和不足，如无法进行有效的故障恢复等。

2. Storm

Storm 是由 Twitter 在 2011 年开源的分布式流数据处理系统。平台采取弱中心化结构，中心节点通过 Zookeeper 分配任务。Storm 把数据处理过程抽象为一个拓扑结构（Topology），把数据流抽象为 stream。通过 Spout 从数据源为拓扑结构提供数据流，由 Bolt 进行数据处理，并在必要时产生新的数据流由下一层 Bolt 继续完成处理过程。与 S4 不同，Storm 自身提供消息处理反馈机制，且能够保障精确处理一次语义。但该平台在故障恢复，系统共存性和启用 Trident 后的并行度方面还存在一些问题。

3. Spark Streaming

Spark Streaming 是由 UC Berkeley 在 2012 年开源的分布式流数据处理系统。系统核心与 Spark^[47] 相同，使用弹性分布式数据集 RDD^[48] 作为处理单元。与其他系统不同，Spark Streaming 在流处理中引入微批次的概念，把数据流分为较粗粒度的数据集，对每个数据集进行 RDD 的转换操作。Spark Streaming 对故障恢复支持较好，但自身也存在一些不足。如微批次的引入增加了数据处理延迟，基于 RDD 的变换难以适用于迭代更新外部状态的

⊖ 现有 Heron 系统相关资料较少，因此暂未列入表 3。

应用场景等。

4. Samza

Samza 是由 LinkedIn 公司在 2013 年开源的分布式流数据处理系统，需要依赖 Kafka^[49]进行数据的获取和传输，其中 Kafka 是由 LinkedIn 在 2010 年开源的分布式队列。Samza 通过 YARN^[50]启动运行，同时也依赖 YARN 的集群监控与故障恢复机制。Samza 本身对状态保存具有一定支持，它提供 LevelDB^[51]作为轻量级的数据存储工具。但对于 Kafka 的高依赖性，使得 Samza 在延迟和数据路由方面存在一定局限，且系统仅能支持至多或至少处理一次语义。

5. MillWheel

MillWheel 是由 Google 在 2013 年公布的分布式流数据处理系统。与其他平台相比，MillWheel 主要强调了对数据严格一次的处理方式，且能够对乱序数据进行有效处理。它采取在基于数据处理单元之间的数据流定义低位线的方式，对数据进行切分和局部排序。平台在数据持久性方面提供了较大的灵活性，允许通过 Bigtable^[52] 和 Spanner^[53] 的备份仓库对状态数据进行持久化。此外，MillWheel 支持动态自适应负载均衡。

6. Heron

Heron 是由 Twitter 在 2015 年公布的分布式流数据处理系统。Heron 与 Storm 的 API 兼容。与 Storm 类似，Heron 把处理流程抽象为拓扑结构（Topology），采用调度器把拓扑结构作为一个由多个容器（cgroup）组成的任务来执行：包括一个 Topology Master，一个 Stream Manager，用于性能监控的 Metrics Manager 和多个 Heron 实例 Spout 和 Bolt。其中调度器可以采用 Aurora 调度器或 YARN、Mesos^[54] 等。系统通过 Zookeeper 保存拓扑元数据。与 Storm 相比，Heron 在系统设计上降低了整体复杂度，有明显的性能提升，更易于调试，有更好的容错性和弹性，更易于在共享设施中部署。

3.2.2 流数据处理系统研究进展

在进行流数据处理系统设计和研究时，通常可以把系统分为不同的研究内容和问题点。根据主要研究问题点，可以分为系统整体结构、数据模型、语义保障、数据存储、计算模型、负载均衡、故障恢复、查询语言等部分。下文将针对这几点研究内容，分别阐述目前系统设计方式及研究进展。

1. 系统整体结构

系统整体结构发展总体经历了由集中式到分布式的变化。早期流数据管理系统主要为集中式的结构设计。如 Aurora 等集中式的流数据管理系统，采用了类似传统数据库系统的模式，产生了如调度器、查询优化器等组成部分。但由于单机内存等诸多限制，集中式的流数据管理系统无法对大量数据进行有效的管理。

随着数据量的增大及对系统处理能力的要求提高，流数据管理系统开始采用分布式的结构设计，由此也产生了一系列新的研究问题。对分布式的系统结构进行再划分，根据系统集群中是否有中心节点以及中心节点在系统中的重要性可以分为中心化，弱中心化以及去中心化结构。

中心化的分布式流数据处理系统中，中心节点作为整个系统的控制中心，掌握着系统的全局信息并负责任务的分配与调度。但正由于中心节点的重要性和唯一性，对于中心化结构的流数据处理系统，存在单点失效的问题（single point of failure）。

弱中心化的系统结构与中心化相比，中心节点的作用较小，仅在系统启动运行时依赖中心节点，系统正常运行后则可以脱离中心节点。如 Storm 中的 nimbus 作为系统的中心节点，主要负责 Job 的提交和分发等工作，系统启动后则可以脱离中心节点运行。但依然存在单点失效的问题，且在系统进行故障恢复时，通常需要中心节点的协助。

去中心化的系统结构则可以有效避免单点失效问题。集群中的所有节点通过可靠的沟通机制进行协调，如 S4 采用 Zookeeper^[55] 进行节点间的互相协调。由于没有主节点对集群负载和故障恢复进行协调，完全依靠节点间互相的通信，对集群监控及故障恢复带来较大难度。

2. 数据模型

数据模型指系统在流数据的处理过程中对数据进行的逻辑抽象，通常为系统内部表示的数据处理单元。

目前大部分分布式系统都把数据流中的每一条单独的数据抽象为一个处理单元，如 S4 把每一条数据抽象为 Event，Storm 将每一条数据抽象为 Tuple。这种处理方式保证了数据处理的时效性，但在某些情况下，过于频繁的数据传递会对整体的处理速率产生负面影响。为此，部分系统引入批次的概念对这种情况进行改善，如 Storm。

Spark Streaming 的数据抽象方式与此不同。它引入微批次（Mini-batch）的概念，主要思想是把数据流切分为不同小段，将每个小段的数据抽象为一个处理单位，即一个微批次。由于 Spark Streaming 与 Spark 采用相同的底层支持，此种抽象方式能够统一流处理和批处理的接口，并依赖 RDD 的变换操作完成对数据的处理。但处理粒度较粗，增加了数据处理的延迟，无法适用于对时效性要求较为严苛的场景。

3. 语义保障

语义保障是系统在对数据处理过程中在处理语义上进行的保障。根据每条数据被完全处理的次数，可以把语义保障的类型分为无保障，至少一次处理，至多一次处理和精确一次处理。其中无保障和至多一次处理语义对系统要求较低，目前主要研究内容集中在至少一次处理和精确一次处理。

由于流数据本身的特性：一经处理，除非进行存储，很难再次获取。为了实现至少一次处理的语义，需要增加对数据存储的支持以保证相应数据可重复获取。数据存储的工作可由流数据管理系统自身支持，或通过采用支持重复获取数据的上层数据源完成。如可以采用 Kafka 作为流数据数据源，由于 Kafka 本身支持对数据的重复获取，从而完成对至少一次处理语义的支持。

精确一次处理语义对于流数据处理系统有较高的要求，需要系统能够记住每条数据的处理状态，并对失败的数据重新处理。Google 在广告系统中使用的 Photon 即支持精确一次处理语义^[56]。目前 Storm 对精确一次处理语义的支持与此相似。

4. 数据存储

流数据处理系统中的数据类型主要分为系统元数据，系统运行中的状态数据，以及流处理的中间结果数据。针对不同类型的数据，系统通常采取不同的管理策略。

系统元数据因其重要性，应保存在可靠的外部存储中。由于元数据数据量较小，目前通常采用分布式协调工具 Zookeeper 进行管理。

系统运行时的状态数据通常由各自节点自行在内存中进行管理。由于内存数据的易失性，相应节点失效后无法进行有效的状态恢复。为此，部分系统引入对状态数据的备份功能以增强对故障恢复的支持。备份数据应由可靠的外部存储负责。而依赖于磁盘的存储方式会对系统整体性能产生影响，无法满足系统对速度的要求。为了在外部存储方面平衡速度和可靠性，相关的研究成果有分布式键值存储 RAMCloud^[57] 和 SILT^[58]；基于内存的可靠分文件系统 Tachyon^[59] 等。MillWheel 即通过外部备份仓库对状态数据进行持久化管理。

流处理的中间结果数据作为流数据的一种形式，往往需要在系统内进行传输。不同系统对此有不同的处理方式。较为直接的方式如节点间建立直接连接，完成中间结果数据的传输。但此种方式需要维持节点间的连接，且一旦节点失效容易导致相应数据丢失。间接的方式如采用分布式队列完成数据传输，Samza 采取的即是此种方式（见图 5）。虽然增加了系统的可靠性，但在一定程度上降低了数据处理效率。

5. 计算模型

目前大部分分布式流数据处理系统采用有向无环图（DAG）作为计算模型。其中图中的有向边代表了数据的流向，点则是相应的变换操作或用户定义的执行单元，但具体形式有所区别：如 Storm（见图 6）、Heron 等分布式流数据处理系统，把流计算过程的有向无环图抽象为一个拓扑结构（Topology），依据拓扑结构进行数据流的传输与处理；而 Spark Streaming 则是把 RDD 的变换过程以有向无环图的形式表示，图中的每个点表示了一种 RDD 的变换操作，并依据有向无环图生成并记录世系（lineage）信息，可用于错误恢复。

6. 负载均衡

在分布式系统中，负载均衡主要任务是保证系统中各个节点的负载相对平均，避免部分节点出现负载过大的情况。影响

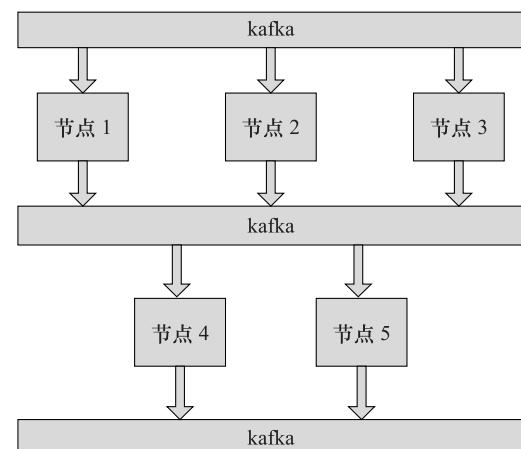


图 5 Samza 通过分布式消息队列存储中间结果

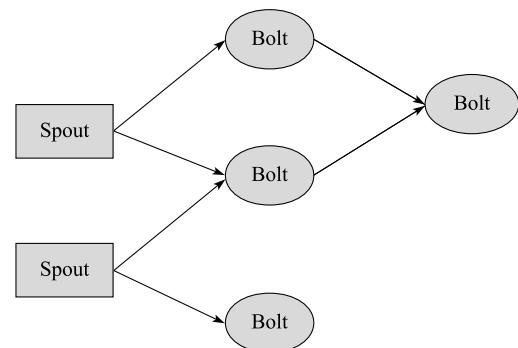


图 6 一种 Storm 流处理的拓扑结构

负载均衡的因素包括数据的划分、路由策略等内容。根据负载均衡策略的确定阶段，可以分为静态策略和动态策略。

静态策略即在系统运行前确定且运行后无法改变。相关研究如文献 [60] 介绍了对高效静态负载策略的制定细节。由于静态策略本身无法在运行时进行修改和调整，无法应对系统运行时出现的负载不均等问题，故具有一定的局限性。

动态策略则可以在运行时根据系统情况进行调整，较适用于现实生产环境，是负载均衡研究的主要内容。Shah 等人^[61]根据对数据划分和路由策略的数据敏感性，把分布式流处理系统中的动态策略分为上下文敏感策略和随机策略。另一方面，根据策略调整时是否需要外界干预又可分为自适应策略和非自适应策略。相关的研究成果如 Exchange^[62]、Flux^[63]等。Storm 自身支持非自适应动态策略，文献 [64] [65] 介绍了基于 Storm 平台的自适应动态调整策略。文献 [66] 采用了 PKG 的分组方式对数据流进行划分，并将此技术应用于 Storm 集群中，系统吞吐量和延迟有了明显的提升，取得了较好的负载均衡效果。

7. 故障恢复

故障恢复是分布式流数据处理系统中的另一重要课题，故障恢复策略的选择决定了系统的容错能力。Hwang 等人^[67]把故障恢复分为精确恢复、回滚恢复和有损恢复三个级别，且在分布式流数据处理系统中设计了一些高可用算法。

目前的故障恢复策略主要通过备份的方式实现。根据备份的方法不同可以分为主动备用、被动备用、上游备份，以及同步检查点等策略。文献 [68] [69] 提出了在分布式流处理系统中的主动备用方法。Storm 则采用了一种使用上游备份的方式进行故障恢复^[40]。Spark Streaming 通过提出 D-Stream 模型^[41]，采用同步检查点的方式建立容错机制，使得流数据处理操作对故障恢复支持更加友好。Gradvohl 等人^[70]对 MillWheel、S4、Spark Streaming 和 Storm 系统的故障恢复问题进行了对比，并对故障恢复策略的组合进行了优缺点的探讨。

此外，Liu 等人^[71]提出了在网络流量分析中的一种容错的流处理架构。Castro 等人^[72]通过使用操作状态管理针对流处理的容错性和扩展性进行整合，以满足云平台对横向扩展和故障恢复的要求。

8. 查询语言

早期的集中式流数据管理系统包含查询语言支持，如 Aurora 支持持续查询语言（CQL）等。目前的分布式流数据处理系统由于自身发展并不成熟，系统对数据的处理操作通常需要用户进行自行编码，自由度较高，系统自身并未提供相应的查询语言支持。已有研究者在现有流数据处理系统基础上，尝试为查询语言提供支持。如 EPFL 实验室的研究成果 Squall^[73]：作为基于 Storm 的查询处理工具，能够实现选择、投影、去重等操作。但其对查询语言支持能力依然存在不足，相关查询优化也较为有限。目前流数据处理系统在查询语言方面的研究发展仍有较大空间。

3.2.3 基于流数据处理系统相关应用研究进展

随着流数据处理技术的不断发展，许多研究者利用流处理技术构建相应的应用，对

不同领域的流数据进行研究分析，取得了许多新的研究进展和成果。

Twitter 作为最为流行和开放的社交网络之一，是现实世界里流数据的重要来源。大量研究者针对 Twitter 流数据进行相关研究，取得了相应的研究成果：Albert 等人^{[74][75]}对 Twitter 实时流数据进行处理，针对用户的实时情绪进行分类，并对情绪的改变进行研究。Hwi-Gang Kim 等人^[76]在 Twitter 实时数据之上，对热点话题进行发现并根据地理位置信息进行聚类。Wang 等人^[77]通过在总统大选期间对 Twitter 数据的实时监控实现了对民众情绪变化的监测与报告。

在广告领域，Google 以分布式文件系统 GFS^[88]为基础，把流数据处理应用于广告系统中，提出了 Photon^[56]，用来分析用户和网站的交互产生的实时事件流数据，并依此进行关联分析。

在流数据上的数据挖掘的相关研究也取得了很多新的成果。Cao 等人^[78]提出一种在能够保障聚类结果准确性的前提下，基于网格的 k-median 聚类算法。Yanai 等人^[79]在 Twitter 实时流数据上进行数据挖掘，能够实时检测出事件照片及实物图片。

3.3 国内研究现状

在国内，流数据相关研究也在不断推进，但尚未出现典型的流数据处理系统。目前国内主要的研究工作围绕在对已有流数据处理系统的改进及流数据环境下的数据处理与分析上。

系统方面，山东大学 Lin 等人^[80]针对现有分布式数据流处理系统的错误恢复和全局存储方面的问题，在对 S4 进行改进的基础上，提出 Pollux 系统。并在此基础上实现了对微博的实时搜索。中国科学院亓开元等人^[81]在实时传感器数据上提出了一种针对高速数据流的大规模数据实时处理方法，提升了对此类数据流处理的实时性和可伸缩性。

国内也在流数据的数据挖掘及分析算法方面开展了相关的研究。山东大学的 Yu 等人^[82]在分布式流处理系统上设计了一种新的索引结构 DSI，并基于 DSI 提出了分布式的 KNN 检索算法 DKNN，加快了 KNN 的计算速度。北京信息工程学院的 Xiong 等人^[83]为了提高查询响应速度，提出了在流数据上基于内存的持续查询索引来实现对查询请求的快速响应。复旦大学的 Duan 等人^[84]在历史流数据上的近似查询上提出了一个新的流 OLAP 框架。

此外，也有把异构计算引入流数据处理中的相关研究。如中国地质大学的 Deng 等人^[85]提出一种新的索引结构 CKDB-tree，通过结合 CPU 与 GPU 计算能力实现数据流上并行动态持续查询的并行处理，但并未支持多个 GPU 和 CPU，具有一定的局限性。

3.4 总结

流数据作为大数据中的一类重要数据，在未来的研究发展中会占据更重要的角色。围绕流数据的管理与分析仍然是一个重要的课题。目前的流数据处理系统在不同的行业

领域逐步得到了应用，但系统本身在许多方面存在不足，如故障恢复、负载均衡、查询语言等。仍有较大的研究和发展空间。此外，在流数据上的数据挖掘和分析等研究工作仍属于起步阶段。具有较高的研究意义和价值。

展望未来流数据处理系统发展方向，应当有如下发展趋势：

1) 提供更加成熟易用的编程接口。目前的流数据处理系统大都只提供简单的编程接口，在易用性等方面仍有欠缺。未来流数据处理系统在提供成熟的编程接口的基础上，可以预定义一些常用的查询模式，提供查询或操作的语言，甚至是提供更为易用的图形化界面等工具。

2) 系统自身稳定性的完善与提升。在系统稳定性与健壮性方面，进行相关研究与改进。提供更好的负载均衡策略和容错机制等。

3) 提供存储机制支持。通过系统内置的存储机制，可以对语义保证等提供支持。

4) 与资源调度器集成。把资源调度工作从流处理系统中剥离，采用 YARN 等资源调度器，完成系统的资源分配与管理等任务。

5) 可调节的数据处理粒度与数据处理延迟。对数据处理粒度和延迟可调可控，使得流处理系统为不同的应用场景提供更好的支持。

6) 对同一数据源的并行高效处理。针对同一数据源的不同处理任务，进行相应的优化，调度等，提升系统性能。

7) 更成熟易用的流处理系统运维监控。目前的许多流处理系统在流处理的运维监控上仍有不足。未来的流处理系统应在目前的运维基础上，针对流处理系统的数据和处理任务，提供更为成熟易用的监控等功能。

8) 流数据处理系统与云计算的结合。如出现公有流计算云等，促进与此有关的如流数据处理系统的安全、多租户、资源分配调度等相关问题的研究。

总之，流数据处理仍面临着许多问题。随着大数据的发展和流数据处理需求的不断增多，势必会对相关的研究提出更多新的挑战，同时也为流数据处理提供了更好的发展机遇和更广阔的发展空间。

4 众包数据管理系统

4.1 引言

关系数据库管理系统自 20 世纪 70 年代被提出以来得到了非常广泛应用，不仅在商业领域，而且在处理诸多其他领域数据，如个人数据、社会数据、科学数据等方面都取得了巨大的成功。然而，随着信息化革命的不断深入，数据采集方式的不断深化，数据呈现“开放世界”的特点——永远不能预见数据的全集和表现形式。然而传统数据库采

用“封闭世界”模型，即把数据库看作数据全集，返回数据库中与查询精确匹配的记录；如果数据库中不包含精确匹配的记录，则不能返回任何结果。大数据时代“开放世界”的特点给以“封闭世界”为基本假设的传统关系数据库管理系统带来了一系列新的严峻挑战，主要体现在以下几方面。

缺乏知识推理功能：传统数据库仅能处理确定性的操作（例如比较大小、相同值聚集等）。然而真实数据非常复杂，例如识别“孙悟空”和“齐天大圣”是否指同一个人，“魔都”与“上海”是否指同一个城市，等等。类似这种“实体识别”问题会伴随当今数据表现形式日趋多样以及数据查询请求日趋多元的特点而日趋突出。显然，传统数据库系统很难处理这类查询，这严重降低了数据库系统的可用性。因此，数据库管理系统十分迫切地需要引入知识推理功能，使数据管理变得越来越智能。

缺乏认知理解能力：数据的类型繁杂性给数据处理带来了认知理解方面的新需求，例如给定一个产品数据库和一些用户评论信息（包含文字评论和图片评论），挑选对这些产品持好评的评论。显然，这类查询依赖于人类的认知理解能力，难以被当前的数据系统所支持。然而，类似的查询在大数据时代非常普遍，例如挑选内容相关的文本，识别满足要求的图片，等等。因此，有必要扩展现有数据库系统的数据处理能力，以应对越来越复杂的数据处理和分析挑战。

缺少错误容忍技术：传统数据库系统假设所采集的数据是完整且准确无误。然而，这一假设在现实生活中并不成立——真实数据库系统中的数据往往存在着各式各样的错误，如不正确、不精确、不完整或者过时陈旧。根据市场研究公司 Gartner 的调研，全球财富 1 000 强公司超过 25% 的关键数据存在不正确或是不精确的现象^[90]。数据质量低下带来的危害也是显而易见的：在美国，数据错误引发的医疗事故每年造成近 10 万患者的死亡^[91]。数据错误产生的经济损失占美国 GDP 的 6%^[92]。在国内，近些年也频发因价格数据错误而引发的事故，如当当网“价格乌龙”^[93]、淘宝网“错价门”^[94]，等等，给商家和消费者带来了诸多负面影响。因此，如何对数据质量进行有效的管理与控制，已经成为当前数据库管理系统的首要问题。

为了使数据库系统成功应对“开放世界”下的新挑战，我们需要引入人类的复杂认知和推理能力。近些年，众包技术迅猛发展，它利用人的认知与推理能力进行计算，成功地解决了很多复杂的问题，为我们带来了一个新的机会：可以通过众包技术来为数据库提供人类的认知推理能力。与传统单纯基于机器的计算技术相比，众包最显著的特点是将复杂计算问题分发给大量的人群进行分布式求解，充分了利用了人的认知与推理能力进行计算。众包在多个领域取得了明显优于机器的效果（例如图片识别、机器翻译、知识图谱构建），成功地解决了很多复杂的问题，体现出了巨大的应用价值。因此，众包为复杂数据管理提供了一条有效途径，能够有效应对数据管理中的新挑战。然而，现有的众包研究都是具体应用具体分析，没有通用的众包数据管理系统来提供统一的平台，因此亟待研究众包数据库来解决目前存在的这些问题：一方面使得传统数据库变得更加智能，另一方面为众包数据管理提供通用的平台。

众包数据库有着广泛的应用前景，不论对国家、政府还是人民都有着非常重要的意

义。首先，众包数据库系统能够为我国闲散劳动力创造更多新就业机会和新就业形式，因此非常适合于我国国情。其次，众包数据库能够响应李克强总理“大众创新、万众创业”的理念，通过凝聚大众的智慧来解决复杂问题，必将创造更多的价值。再次，众包数据库不论在国内还是国际都处于探索阶段，如果我们能够在此时加大力度，开展众包数据库管理系统的关键技术研究，我国很有可能在这一新兴而重要的领域占据国际领先地位，有机会打造出在新时期数据管理行业的属于中国的 Oracle 和 Google，从根本上改写全球数据管理市场的版图。

4.2 众包数据管理

4.2.1 众包基本概念和工作流程

自 2006 年 Howe 首次提出众包的概念以来，学术界对众包的定义一直没有一个公认的结论。文献 [128] 总结了 40 种不同的众包定义，这些定义从不同的角度对众包进行了描述，对比这些定义可以得到众包的基本特征：

- 1) 采用公开的方式召集互联网大众。
- 2) 众包任务通常是计算机单独很难处理的问题。
- 3) 大众通过协作或独立的方式完成任务。
- 4) 众包是一种分布式的问题解决机制。

根据这些特征，概括来说众包是一种公开面向互联网大众的分布式的问题解决机制，它通过整合计算机和互联网上未知的大众来完成计算机单独难以完成的任务。

众包的主要参与者包括任务请求人 (requester) 和工人 (worker)，也称为任务完成人。他们通过任务 (tasks) 联系到一起。

当任务请求人打算利用众包完成自己的任务时，需要按照以下步骤来使用众包。

- 1) 设计任务；2) 利用众包平台发布任务，等待答案；3) 拒绝或者接收工人的答案；
4) 根据工人的答案整理结果，完成自己的任务。而工人使用众包的主要步骤包括：
1) 查找感兴趣的任务；2) 接收任务；3) 回答任务；4) 提交答案。

从时间维度来考虑，可以把众包的工作流程分成三个阶段：任务准备、任务执行和任务答案整合。其中任务准备阶段包括：任务请求人设计任务、发布任务，工人选择任务；任务执行阶段包括：工人接收任务、解答任务、提交答案；任务答案整合阶段包括：请求人接收/拒绝答案、整合答案。

众包任务按照其复杂性、难度和粒度分为复杂任务 (complex task)、简单任务 (simple task)、宏观任务 (macro task) 和微观任务 (micro task)。顾名思义，复杂任务主要是解决一些非常复杂的、不能轻易完成的任务，例如编写一个软件，开发一个网站等。简单任务主要是解决一个较容易完成的任务，例如设计一个 logo、翻译一段文字等。宏观任务是指一个人较容易完成的任务，宏观任务的粒度较大，例如写一个饭店的评论。而微观任务则粒度更小，例如对一个图像打标签、查找某个酒店的电话号码、录入一张

名片、判断两条记录是否为同一个实体等。

4.2.2 众包平台

众包任务的发布和答案的收集是通过众包平台来完成的。众包平台主要分为两大类：一类是商用的众包平台；另一类是社交网络、论坛等社交平台。

目前有很多商用的众包平台，其中主流的众包平台包括：Amazon Mechanical Turk (Mturk)、CrowdFlower^①、samasource^②、CloudCrowd^③等。国内也有很多众包平台，例如清华大学开发的众宝^④、猪八戒^⑤、三打哈^⑥等。商用的众包平台根据任务请求人和工人的不同需求提供相应的服务，并向任务请求人收取一定的管理费用。

在主流的商用众包平台上发布的任务主要是微观任务，每个任务叫做 Micro Task（微任务），其主要原因包括：

- 1) 工人更喜欢完成一些粒度比较小的任务，这样可以在很短时间内完成一个任务。而复杂任务很难在短时间内完成，需要花费工人很长时间。由于工人一般利用较短的空闲时间来完成任务，因此工人通常不会挑选复杂任务。
- 2) 任务请求人很难衡量复杂任务的价格，因此不利于支付报酬。
- 3) 微观任务具有更好的交互性，任务请求人可以多次和工人交互获得更加准确的答案，此外任务请求人还可以通过众包来多次尝试自己新的想法。

4.2.3 众包数据管理基本查询算子

在众包算子方面，研究者将人群能够提供的数据操作抽象成算子，例如：选择/过滤算子^[101-103]、连接算子^[104]、填充算子^[105]、排序及求解 Max 算子^[104,106,124]、统计分析算子^[107]。众包算子和传统算子的区别在于，众包处理时可能产生错误结果，因此我们需要容忍这些错误结果，并从中推断出正确结果。

在^[101]中描述的过滤算子中，为了降低期望花费以及减少过滤操作的期望误差，作者提出了一种有效的剪枝算法，将冗余的问题去除掉。然而^[101]中并未考虑不同的工人的差异性以及可能获得的先验的信息，在考虑以上两个因素的情况下^[103]中进一步提出了两种过滤操作的优化算法。文献^[102]研究了基于众包的选择操作，作者将花费和延迟作为主要的优化目标，提出了多种启发式算法。

文献^[104]是最早的基于众包的连接算子的研究工作之一，针对连接操作给出了三种方法，一种是简单的比较方法，一个任务只包含一对元素；第二种是简单的批量方法，一个任务包含多对元素；第三种是智能的批量方法，任务不再是元素对的形式，而是一组单独的元素，工人需要在这些元素中查找匹配的元素。对于排序操作，他们提出了基

① [http://crowdflower.com/。](http://crowdflower.com/)
② [http://samasource.org/。](http://samasource.org/)
③ [http://www.chinacrowds.com/。](http://www.chinacrowds.com/)
④ [http://www.naoliku.com/。](http://www.naoliku.com/)
⑤ [http://www.zhubajie.com/。](http://www.zhubajie.com/)
⑥ [http://www.sandaha.com/。](http://www.sandaha.com/)

于比较和基于打分的任务提问方法。

文献 [105] 描述了斯坦福大学开发的一个基于众包的结构化数据收集系统，作者将结构化的数据表格呈现到不同人的面前，利用人的经验和知识完成结构化数据的收集，该系统能够实时的收集不同的工人的答案并且能够利用工人的投票来解决潜在的答案不一致性问题。

文献 [106] 研究了基于比较的 Max 算子，作者将收集自大众的成对的比较结果归结到一个有向无环图上，在此图上提出了 4 种启发式的算法来求解最好的元素，此外，作者在文中同时设计了 4 种选择问题的启发式算法。文献 [124] 是斯坦福在众包环境中的另一篇关于 Max 的工作，作者提出了两种求解 Max 的算法，一种是基于冒泡排序的算法，另一种是基于锦标赛的算法。综合实验来看，基于锦标赛的算法取得了更好的效果。

文献 [107] 研究了众包数据库中的 count 操作，作者提出了两种收集数据的方式，一种是基于标注的方式，另一种是基于统计的方式，来得到大众对数据集中满足特定条件的条目一个估计。

4.2.4 众包查询优化

众包查询优化一般研究计算延迟、结果质量、金钱开销三者之间的平衡。这里我们介绍如何降低金钱开销，即减少问题数。^[118,122] 研究了对 Max 操作的优化策略。^[118,121] 对众包中的 Top-k 操作进行了进一步的优化。^[120,123] 探究了排序操作的优化方法。^[118,120,121] 基于成对的比较来进行排序或者求解 Top-k。^[122,123] 通过引入打分操作，在工人质量较高的情景下能取得比单纯的比赛更好的效果。下面我们详细介绍上述的几种查询优化算法。

在^[118]中，作者提出了针对众包环境中 Max 和 Top-k 操作的优化，在 Max 操作中，作者采用了一种两阶段的算法，在考虑不同的工人的质量差异的情况下提高算法的鲁棒性。在此 Max 算法的基础上，作者又提出了众包环境中 Top-k 操作的优化，第一步通过分桶的方法得到 Top-k 的候选集，然后在此结果的基础上利用一种基于堆的算法来计算最终的结果。文献 [122] 提出了一种针对众包环境中 Max 操作的优化算法，作者提出了一种两阶段的策略来降低整体的开销以及提升任务完成的质量，在第一个阶段中，作者利用打分的操作去除一部分候选的答案，在第二个阶段中利用成对的比较的方式，通过极大似然估计的方法来得到一个最优的结果。然而极大似然估计的效率非常低，所以作者在文中同时设计了一种改进的 PageRank 的算法来集成两个阶段所收集到的打分和比较的答案。

文献 [121] 提出了一种对 Top-k 的优化算法，作者设计了一种迭代的算法，在每一轮迭代中，计算出很大概率在 Top-k 中的元素以及很可能不在 Top-k 中的元素，在选择问题的时候避开这些元素来减少整体的花费。

文献 [120] 对众包中的排序操作做了显著的改进和优化。首先作者对不同的工人的质量和得到的成对的问题的回答进行了基于 Bradley-Terry 模型的建模，其次作者设计了一种基于主动学习的方法来选择每一次所问的问题来提升整体的排序的质量。文献 [123] 和 [122] 类似，也是通过引入打分的操作来提高任务的质量，作者同样采用主

动学习的方法在每一轮的迭代中选择信息量最大的进行打分或者比较操作，显著提高了排序的质量。

4.2.5 众包质量控制

在工人提交答案后，可以通过各种算法来保证结果的质量。最简单的办法是把一个任务分配给多个（奇数）工人来完成，然后通过多数投票原则（少数服从多数）来获取最终结果。由于多数投票方法是假定每个工人的答题准确率是一致的，没有考虑工人的多样性，而通常不同工人的答题准确率差异较大，比如，欺诈者的答题准确率较低，因此采用这种方法得到的最终结果往往不够准确。

针对这一问题，一些研究做出了改进，将工人的答题准确率运用到结果的估计中，使得最终结果的质量有了很大的提高。Liu 等人^[100]通过增加测试题目得到工人的答题准确率，利用贝叶斯理论将工人的答题准确率和工人给出的答案结合起来得到最终的结果。文献 [125] 提出了一种概率模型，该模型基于因子图，通过综合工人答案、工人的答题准确率等因素得到结果。这两种方法都是区分了不同工人的答题准确率，但是都假定了工人的答题准确率是固定的，即：在完成任务的过程中，工人的答题准确率保持不变。然而，随着时间的变化，工人的答题准确率通常是变化的，比如，工人在完成任务过程中，随着对任务了解的增多，他的答题准确率会越来越高。文献 [109, 126] 提出了一种反映工人答题准确率变化的方法。这种方法是基于 EM (Expectation-Maximization) 算法得到的最终结果，通过混淆矩阵 (Confusion Matrix) 来反映工人的答题准确率。EM 算法需要两个步骤进行迭代计算，直到算法收敛。第一个步骤是利用已有的工人答题准确率估计值，对所有的问题分别进行计算，得到每个问题结果的估计值；第二步是利用第一步得到的结果来计算每个工人的答题准确率。由于 EM 算法每次迭代都要重新计算每个问题的结果和每个工人的答题准确率，因此当问题数目较多或参与回答问题的工人数量较多的时候，EM 算法的运行时间较长，代价较大。以上计算工人答题准确率的方法都难以实现准确性与实时性的平衡，文献 [127] 提出了一种新的工人模型，通过该模型可以及时准确地得到工人答题准确率，他们利用工人每次返回的新答案，结合答题准确率，设计了两种增量式的策略来推断最终的任务结果，进而实现高效准确地得到任务结果。

此外，我们也可以利用工人的答题准确率来进一步提高多数投票原则得到的结果质量，其基本思想就是根据工人的答题准确率来对每个工人进行打分。回答准确率越高的工人赋予权重越大，相反则权重较小。最后通过考虑权重进行加权评价工人提供的答案，根据加权分值确定最终结果。

4.3 国外研究现状

近年来，国内外一些著名的研究机构均开展了将众包融入数据管理的研究项目，取得了令人瞩目的成绩。下面从众包系统开发、众包算子设计、众包质量控制、众包应用研究等方面对现有工作及其存在的局限性进行总结与分析。

在系统开发方面，美国加州大学伯克利分校与瑞士苏黎世理工学院的研究者通过扩展传统结构化数据查询语言 SQL 来支持基于众包的数据操作，开发了众包数据库系统 CrowdDB^[97]， CrowdDB 实现了和传统 SQL 查询语言的兼容，应用开发者可以书写 SQL 语句而不需要考虑哪些操作在机器上完成哪些操作利用众包来完成。 CrowdDB 中包含着不同任务的模板，能够根据不同的查询的需要自动生成相应的网页代码并发布到 Mturk 平台上。

美国麻省理工学院的研究者将 SQL 语言和用户自定义的众包操作结合起来，开发了众包数据管理系统 Quirk^[98]， Quirk 系统中包含一个任务的缓冲池，只有当一个任务无法从任务缓冲池中直接得到答案或者间接推导得到答案时， Quirk 系统才会根据查询的需求利用任务编译器生成特定的任务发布到 Mturk 上面。

斯坦福大学的研究者对众包数据模型进行重新定义，提出了声明型众包数据库系统 Deco^[99]。 Deco 中不仅定义了传统的关系模型，而且引入了提取规则和解析规则。开发人员通过定义自己的提取规则从大众中获取数据，然后利用自己指定的解析规则获得最终的结果。

新加坡国立大学的研究者设计并研究了众包算子的查询优化，开发了众包数据分析系统 CDAS^[100]。 CDAS 主要由 3 层构成。首先是平台层，联系任务发布者和用户；其次是质量控制层，通过合适的定价以及挑选合适的工人来提高任务完成的质量；最后是应用层，应用层将一个复杂的任务分解成若干微任务，通过平台层发布出去。

不过，现有系统大多基于传统的关系模型进行扩展，并未很好地解决前面提到的众包数据管理中的“开放世界”、数据质量控制，以及质量感知的查询优化等问题；此外，现有系统在人机混合处理及优化方面尚存在局限性，还没有形成有效的集成人群认知推理能力和机器大规模计算能力的理论框架与高效算法。

在算子设计方面，斯坦福的研究人员先后发表了 3 篇关于 Max 算子的文献 [106, 122, 124]，文献 [106, 124] 主要提出了几种基于比较的启发式算法来求解 Max。文献 [122] 引入了打分的操作来做进一步的优化。众包中过滤操作也来自斯坦福研究人员的工作^[101,103]，这两篇论文集中解决了众包环境中过滤操作的优化问题。此外文献 [102] 尝试解决选择算子的设计，^[104] 初步探究了排序和连接操作的设计，^[105] 开发了一套完整的结构化数据收集系统，^[107] 提出了两种进行统计分析问问题的方式。综合多个众包算子设计的工作来看，现有研究对计算过程中成本、准确率以及完成时间之间的权衡关系进行的研究还比较初步，一些基本的问题还有待进一步的探讨，例如：定价机制如何同时影响准确率、完成时间和金钱开销，能否通过调节价格对算子做进一步的优化；如何在资源受限的情况下（如金钱受到预算的制约）下，进行多算子整体的查询优化，等等。这些问题都有待深入研究。

在众包质量控制方面，现有的工作大多采用多人分配任务和投票聚合答案的策略，例如多数投票策略。为了进一步提高效果，它们对答题人群的准确率进行了估计，方法主要分为两类：第一类是通过混入少量带有标准答案的题目估计答题人群准确率^[100,108]；另一类是通过 EM 算法同时对准确率和聚合结果进行估计^[109]。现有工作对于质量控制的

很多重要问题都还没有深入探讨，例如前面提到的构建面向众包群体行为的可靠性模型，描述并预测用户在完成不同主题任务时可能存在的准确率差异；分析定价模型对于众包质量的影响；如何实时评价用户质量；以及如何在资源受限情况下构建高效的质量控制策略等。

在众包应用研究方面，CrowdSearch^[110]将众包应用于图像搜索中，较为准确地对图像相关性做出了估计；Solyent^[111]是一个众包文本处理系统；CrowdER^[95]提出了一种人机混合的实体识别技术，将群体智慧与机器处理融合起来解决数据质量问题。此外，众包也被广泛地应用于一些具体的数据处理过程中，如模式匹配^[96]、图数据搜索^[112]，等等。然而，现有的众包应用场景研究还相对简单，缺乏对复杂应用场景的深入研究。例如，在构建知识图谱的过程中，如何发挥众包认知推理作用，控制众包的金钱开销与完成时间，有效地与大数据处理技术融合起来，得到更好的结果。此外，目前众包应用都根据不同应用设计不同算法，缺乏通用的平台来支持众包数据处理。

4.4 国内研究现状

清华大学数据库研究组开发了名为“众宝”的基于众包的数据收集和分析系统。首先，“众宝”系统提供了丰富的模板，方便用户设计多种形式的问题。其次，“众宝”系统具有丰富的 API，使得任务发布者能够设计策略主动选择自己需要的工人。另外，“众宝”系统有着功能完备的手机客户端，能够结合工人的地理位置信息，方便任务发布者指定回答问题的工人的地理范围。该组研究了基于众包的实体连接技术^[95,119]以及迭代式众包计算方法^[117]，此外我们还开发了一个质量感知的基于 Mturk 的任务分配系统^[116]。

清华大学刘云浩教授提出了群智感知计算，通过发挥“人多力量大”的特点，将大量草根用户拧成一股绳，形成随时随地、无孔不入、与人们生活密切相关的感知系统^[113]。哈尔滨工业大学的刘挺教授通过众包游戏软件构建了中文的语义相关性词典^[114]。哈尔滨工程大学的张志强教授提出了阶段式动态众包质量控制策略^[115]。

但是总体来讲，国内针对众包的研究工作还不多，缺少对众包中任务管理、数据管理、质量管理等多方面的研究工作，缺少资源受限的众包计算方法，缺乏高通用性的众包平台和应用系统。

4.5 总结

众包数据库管理在国内尚未得到应有的重视，在国际上也尚处于探索阶段，还存在着很多重要问题没有解决。

众包数据管理的基础理论有待探索：众包数据管理无论是在模型还是在访问方式上都与传统大有不同。一方面，与传统数据库系统的“封闭世界”假设不同，众包数据属于“开放世界”，永远不能预见数据的全集和表现形式。基本假设的不同会给数据模型带来根本的差异；另一方面，众包数据管理需要综合利用人群和机器，进行人机混合操

作，这会给数据库的访问方式带来很大的变化。针对这些显著变化，亟待我们对众包数据库管理的基础理论进行深入研究，从而对众包数据库系统及应用的开发起到指导作用。

众包数据的质量控制机制研究不足：由于众包将任务分配给互联网上未知的人群，而且人群回答的质量参差不齐，因此收集的结果可能是不精确或是存在错误的。例如：在众包任务分配的过程中，可能会遇到欺诈者或是恶意攻击者提供错误的答案，从而降低收集数据的质量。尽管现有工作对众包质量控制进行了一定程度的探讨，但尚有很多基础问题有待研究，例如：如何对众包群体在不同主题任务中存在的质量差异进行建模；如何更好地分配任务提高质量控制的效果；如何充分利用人群的认知和推理能力，等等。

众包数据库的查询优化技术尚需提高：众包数据库的查询优化远比传统查询优化复杂：一方面需要考虑如何设计众包问题，在最少的成本下、最短的时间内，收集到质量尽可能高的结果；另一方面，需要设计通用的人机查询优化机制，探讨如何将复杂的任务分解为人群计算算子和机器计算算子，如何决定算子的执行顺序，以及如何集成两类算子进行整体的查询优化。以上这些问题目前都没有得到深入研究。

通用的众包数据库管理系统还未形成：当前很多众包数据管理都局限在特定的应用中——针对不同的应用开发不同的众包处理系统，尚未形成较为公认的通用众包数据库管理系统，能够有效地处理之前提到的质量控制、查询优化等一般性问题。这为我们提供了一个良好的机会：如果能够加大力度进行通用众包数据库管理系统研究与开发，很有可能重现 20 世纪 IBM、Oracle 在关系数据库上的荣光。

因此，我们认为当前在国内开展众包数据库研究是必要的，也是刻不容缓的。如果我们能够在此时加大力度，开展众包数据库管理系统的关键技术研究，我国很有可能在这一新兴而重要的领域占据国际领先地位，有机会打造出在新时期数据管理行业的属于中国的 Oracle 和 Google，从根本上改写全球数据管理市场的版图。此外，众包数据库系统能够是为我国闲散劳动力创造新的就业机会和形式，无论对国家、政府还是人民都有非常重要的意义。

5 在线数据分析与管理系统

5.1 引言

随着各种应用中数据量的快速增长，用户查询和分析任务复杂度的增加，数据维度的增长，快速高效的返回用户查询结果越来越成为大数据系统中的一个重要挑战。在传统的结构化数据处理中，关系型数据库系统利用数据结构化的特点，发展总结出了一套成熟的数据索引，查询优化，以及查询执行的技术，基本上能够达到有效的优化查询计划，减小查询响应时间的系统要求。但是在大数据的背景下，这套技术很难被直接利用，

面临着很大的挑战。一是大量数据一般都需要一个分布式集群进行分布式存储，以及分布并行的计算。这使得传统的单机查询优化技术很难在集群上直接应用。二是大数据经常是非结构化的，或者是异构的，传统的针对结构化数据所设计的数据索引和查询执行技术就不能有效地发挥作用。三是用户查询和分析需求不能用简单的语义逻辑来代表，这样就使得系统经常需要处理 ad-hoc 查询或者复杂的嵌套式查询，传统的查询优化方案很难有效处理这类查询。

基于以上观察，目前学术界和工业界的一个研发重点就是如何在大数据系统中支持实时在线的数据查询。这类系统主要关注两个方面：一是交互性，还有一个是实时性。交互性体现在处理用户查询过程中系统及时不断地提供反馈，这样使得用户能够快速地做出反应和根据反馈结果更改或优化下一步的查询条件，来找到最相关和最有意思的结果。实时性是指系统能够快速响应用户的各类不同查询输入，在尽可能短的时间内提供查询结果。

在线数据分析与管理系统在很多领域都有着广泛的应用，譬如用户去网上购买商品，常常会根据某些属性的组合来检索合适的商品，如果系统反馈的时间过长，用户很可能失去耐心，不再购买。这类的查询在电子商务、电信、金融等诸多领域都有着非常广泛的应用。

实现在线数据分析与管理系统大体上有两种思路：一种思路是在分布式的环境下通过增加硬件资源（CPU、内存等）来减少数据查询的时间；在这种思路的指导下，Google 开发出了 Dremel 系统，主要用于 Web 数据的分析与查询。受到 Dremel 的启发，很多企业开始研发类似的产品，目前比较具有代表性的就是 Cloudera 公司的 Impala 系统。在 Dremel 之外，Apache 社区开发了 Apache Drill 系统，该系统和 Dremel 的应用场景略有不同，但是都能在极短的时间内对海量的数据进行查询和分析。另一种思路则在数据精度和查询时间的延迟上进行了权衡，以牺牲一定的数据精度来换取查询时间上的大幅度减少。这类系统的典型代表有 UCB 的 AMP 实验室开发的 BlinkDB 等，以及数据库领域经典的在线聚集技术。

本报告会对在线数据分析与管理系统的主要研究问题以及国内外研究现状进行介绍，并对未来的发展趋势进行展望。

5.2 主要研究问题

实际的应用对于实时在线分析的需求越来越强烈，但是传统的数据库管理系统很难应对大数据环境下的在线分析。目前针对在线分析的研究，主要集中在处理模型、数据采样等问题。

5.2.1 在线分析的处理模型

针对大数据，主要的处理模型无非就是精确数据分析和近似数据分析。一般来说，用户还是希望得到精确的数据分析结果。但是面对如此海量的数据，很难设计

出一个通用的且能实时获取精确结果的在线分析系统。针对某些特定的应用场景，的确可以设计出返回精确结果的系统。但对于更多的应用场景而言，在有限时间内返回近似结果可能是一个更现实的选择。对于近似数据分析而言，目前比较常见技术包括在线聚集等。针对具体的应用，如何选择合适的处理模型和相关技术会是一个重要的问题。

5.2.2 数据采样

数据采样是大数据在线分析的一个重要问题。如果采用近似的数据处理方法，那么将在很大程度上依赖数据采样。即使是精确的在线分析，在很多场景下，通过对小数据进行采样，分析出整体分布的一些特性，可以有效提升后续的分析效率。常见的概率采样包括简单随机采样、系统采样、分层采样、整群采样和多阶段采样等方法。一般来说，最常用的是简单随机采样。在很多场景中，这种简单的采样方法可以取得很好的效果。但是对于某些复杂的分析来说，这种采样方法无法取得好的效果。但如果采用非常复杂的采样方法，处理效率有可能会大幅下降。如何针对应用场景，设计新颖和合适的采样方法会是大数据在线分析的一个重要研究问题。

5.2.3 查询处理与优化

查询处理是连接用户请求与各种统计计算的重要桥梁，而查询优化则可以进一步提高处理速度。针对各种场景的查询处理和优化技术，工业界和学术界都进行了一定程度的研究。从支持的查询接口和查询语言来看，早期的一些系统如：Bigtable、HBase 和 Cassandra 等仅支持一些基本的数据插入和获取接口。随后很多公司和研究机构在丰富查询语句上开展了工作并提出一些“类-SQL 语言”，例如 Yahoo! 的 PigLatin，Facebook 的 HQL，微软的 SCOPE 和 DryadLINQ，以及 IBM 的 Jaql，等等。但是目前这些研究工作主要集中在相对简单的查询处理和优化上。对于经常使用的一些复杂操作，比如 join 等，还有很多亟待解决的问题。

5.2.4 复杂操作的支持

数据分析是一个很复杂的问题。目前的大数据在线分析主要面临两个部分的问题：一方面对于类似在线聚集的近似计算方法，如何全面支持各种操作符是一个具有挑战性的问题。当前的在线聚集主要支持 Count、Sum 等几个简单的 SQL 查询。对于 Max、Min、Top-K 等更加复杂的查询操作目前还没有很好的支持。另一方面更加复杂的分析操作，比如聚类、矩阵分解等尚未有好的解决方案。

5.3 国外研究现状

针对在线分析的特点和相关数据需求，国外的研究机构和企业主要从实际的系统出发，构建出若干能够满足特定场景下大数据在线分析与管理系统。这些实际的系统能够

集中体现相关的在线分析技术。本节主要介绍其中几个具有代表性的系统。

1. Dremel 系统

Dremel 系统是 Google 公司针对海量 Web 数据实时查询需求而开发的一套系统。Dremel 是面向大规模集群而设计开发的，所以它可以在成千上万台节点上进行快速高效的并行查询。根据 Google 的技术白皮书，Dremel 系统可以在 10 到几十秒内就完成对无索引的 350 亿行数据的扫描。Dremel 的两个核心设计思想是列存储和树形的查询处理架构^[129]。

1) Dremel 内核对数据进行按列存储的方式进行处理，利用压缩技术可以大大减少数据吞吐量，并带来扫描数据效率的极大提高。

2) Dremel 的另外一个核心思想是树形体系架构来处理查询。这样的好处是可以提高查询处理中的数据 locality，快速高效进行相近数据的局部聚合，减少数据吞吐量，大大提高在上千台节点间查询任务分配和数据传输的效率。

具体来说，Dremel 的列存储技术将一个数据记录分按列分成多个数据值，并将每个数据值存储在不同的存储媒介上。而传统的数据库通常将一个整数据记录存储在一个存储媒介上。这样的存储方式意味着 Dremel 可以大大减少数据访问量，例如“Select top (title) from foo”就可以避免访问每个记录的整条内容，而只用提取 title 这一列。这样的存储方式的另外一大优点就是高效能的数据压缩。由于按列存储带来了相同类型的值大块存储，对于提高压缩性能有很大帮助（尤其当列值值域小的情况下）。压缩性能的提高意味着系统所需要的吞吐数据会大幅减少。

列存储的一个劣势是不能有效地支持数据更新（update）。针对这个问题，在 Dremel 系统中，对已有数据的更新是不支持的。也就是说，Dremel 系统的主要应用场景是 OLAP/BI 等只读数据处理系统。列存储在传统的数据处理系统中已被广泛应用，但是 Dremel 是第一个将这项技术成功应用到上千甚至上万台集群的数据处理系统。

在设计和研发 Dremel 系统中的一个主要挑战是如何将用户查询分成小的模块并将其分配到成千上万台的计算节点中去，并快速高效地回收、聚合查询结果。Dremel 的答案是设计并使用一套树形的查询执行系统架构。这个架构使得 Dremel 可以将一个用户查询分区成一个可以大规模并行计算的分布式查询树，并且将一个查询推到树形结构底端进行分布式并行处理，然后在聚合这些查询结果。

最后，Dremel 设计并使用了一套 strongly-typed 嵌套式数据模型。具体而言就是每个记录属性都有一个指定的数据类型，数据记录的机构可以是嵌套型的（与流行的 JSON 记录类似）。为了便于使用，Dremel 开发了一套类型与 SQL 的系统查询语言，这样用户可以轻松快捷地学习并开始使用 Dremel 系统。Google 将 Dremel 系统在其云平台上通过 Google BigQuery 系统开放给开发者和普通用户使用，凡是使用 Google 云平台的开发者和用户都可以通过 BigQuery 界面或者 API 来调用 Dremel 内核进行大集群上的互动式数据分析。

2. Apache Drill 系统

由于 Google Dremel 系统并没有开源，Apache Drill 系统是 Apache 社区基于 Dremel 的

论文^[129]而设计开发的一套开源系统^[130]。其基本设计理念与 Dremel 有很多相同之处。Drill 的一个核心设计目标就是能够拓展到上万台节点并且能够在数秒内处理 petabytes 数据和数十亿条记录。Drill 也是支持类似 SQL 的查询语言和界面，主要用来支撑和处理在大规模分布式存储的结构化和半结构化数据上的互动式的数据分析。

Drill 支持多个不同的数据模型，具体而言可以分为两大类：有数据模式的，例如 Apache Avro 和 Protocol Buffers；无数据模式的，例如 JSON、BSON，等等。Drill 的查询语言是 DrQL，一种类似于 SQL 的但是支持嵌套型数据的语言。DrQL 的一个特点是和 Google 的 BigQuery 系统兼容，另一个特点是能够高效地支持列存储数据，尤其是在查询过程中避免了从列存储模式到行存储模式的数据重组。除了 DrQL 以外，Drill 也支持类似 MongoDB 里面的 Mongo 查询语言，这样可以有效地处理类似 JSON 一样的数据模型。

Drill 的存储引擎是一般的 DFS（分布式文件系统），例如 HDFS。所以 Drill 一般可以部署在一个 Hadoop 集群上面。与 Dremel 的一个明显不同是，Drill 可以同时支持列存储处理和行存储处理。这样的架构对支持不同的数据类型是很有必要的。如果系统需要，Drill 也可以在系统内部进行数据格式的转换，将行存储变成列存储之后再进行处理。这样的体系架构使得 Drill 能够非常灵活地导入并在一个查询中同时处理多个不同数据源的数据。更灵活的是，数据的模式可以在不同查询之间变化，甚至是在同一查询的处理过程中就发生变化。

Drill 的查询优化流程可分为 rule-based 和 cost-based 优化。这点与传统的关系型数据库系统中的查询优化流程并没有太大的不同，但是具体的优化方法就有很多的新技术点，这里就不一一展开了。

3. Cloudera Impala 系统

Impala^[134]是 Cloudera 公司开发的基于 MapReduce/Hadoop 集群的互动式数据分析系统。与 Dremel 和 Drill 类似，Impala 也是以支持 SQL 查询为目标，其支持的数据源可以是 HDFS，也可以是 Hive 里面的数据仓库，还可以是基于 HBase 的 Key-Value Store。Impala 系统的一些特性和优点如下。

- 利用本地节点数据处理来减少集群中数据传输开销。
- 统一的 metadata 管理引擎，简化了 metadata 的管理，存储和使用。
- 由于 Impala 支持多种不同数据源和数据接口，系统避免了高开销的数据转换。
- 所有的数据一旦引入到 Impala 系统中，用户即可开始查询，不必要采用 ETL 流程来清洗整合数据。
- 优化的资源调配体系可以最大化地利用集群中的所有硬件资源。
- 单一的资源池简化并提高了系统的整体可拓展性。

Impala 系统的另外一大特点是能够同时支撑多种不同的数据分析需求场景。它既可以作为一个互动式的数据分析引擎，也可以作为线下的批处理引擎，同时也支持关键词为主的互动式搜索和分析。Impala 系统的查询执行和优化模块的核心思想与 Dremel 和 Drill 并无太大区别。具体来说就是尽可能地利用本地节点数据的 locality 来最大化本地节点数据聚合，同时减少网络数据传量。每个节点都包含 Query Planner、Query Coordinator

和 Query Executor。其中 Query Coordinator 会根据 Query Planner 的输出来调动本地的 Query Executor 以及相关的其他节点的 Query Executors 来完成一个查询。本地聚合结果会按照数据流的方式传输给其他的相邻节点的 Query Executor。对于 Join 操作，与传统数据库类似，系统只会考虑 left-deep tree 计划。在需要调用多个节点进行分布式查询时，对于 join 操作，系统考虑两种方案：分别是 Broadcast Join 和 Partitioned join，这点与其他类似系统并无大的区别，例如 Spark SQL^[135]。

4. BlinkDB 系统

当数据量过大或者用户查询反应时间要求更低时，支持互动式在线数据查询的另外一个思路就是利用数据采样技术。早在 1997 年的 SIGMOD 大会上，学术界就提出了在线聚合的概念，其具体思想就是利用随机样本来获得小的数据量并根据很小的数据量来估算最重的聚合结果。随着随机样本的不断增加，聚合估算的结果也越来越精确^[133]。

核心挑战有两个：一是如何快速高效地生成满足查询条件的随机样本，二是如何根据随机样本对不同的查询进行结果估算并对结果的质量提供精度保证。BlinkDB 是基于以上思想开发的一套系统^[131,132]。对于第一个挑战，BlinkDB 的基本解决方案是提前对数据采样，并将样本保存在内存中。对于不同的查询条件，对样本进行过滤来生成满足当前查询条件的样本。具体来说，BlinkDB 又支持 Uniform 和 Stratified 的两种样本模式，其中 Stratified 样本模式对于支持 GroupBy 等分类查询条件尤其有效。同时为了减少内存中所保存的样本数量，BlinkDB 会对历史查询模板进行分析，从而找到最常见的查询条件模板并对这些查询模板生成样本。以上解决方案的优势在于简单有效，非常方便与已有系统结合。例如 BlinkDB 就在 Spark 系统上进行了实现。但是这个解决方案的一个缺点是需要提前生成并保存大量样本（因为实时生成样本的开销过大），并且当多个查询重复使用样本的时候，样本独立性没有得到有效的保证。

第二个挑战是如何对不同的查询条件生成估值，并对估值质量提供精度保证。在这一点上，BlinkDB 系统基本上借鉴了经典的在线聚合技术里面的结果，针对常见的聚合查询（例如，求和、平均等）提供了估值公式和精度估算公式。

5.4 国内研究现状

随着国内互联网产业以及传统行业对互联网 + 技术升级的产业需求，互动式实时数据处理的需求也与日俱增。国内的学术界和工业界在这个方向都投入了很大的研发力量。限于篇幅所限，这里重点介绍在系统和关键技术等方面的一些近期发展。

国内的互联网企业，比如 BAT 巨头，很早就面临着如何在大数据环境下快速高效地进行数据分析的需求和挑战。一些传统企业在发展过程中也慢慢地遇到了类似的挑战，比如华为在处理运营商数据的时候就面临着如何满足运营商对数据分析响应时间的要求，针对移动数据量的几何式增长的解决方案越来越需要互动式实时性技术的支持。这些实际需求推动了这些公司在互动式实时性数据分析技术上的投入和所带来的相关技术与系统的快速发展。比如华为公司就推出了面对大数据分析的 FusionInsight 平台，其中一个

关键解决方案就是支持在大数据上面的互动式和实时性的数据分析与整合，以及建模和数据探索。阿里巴巴也投入了大量的资源开发了并行分布式数据库与处理系统 Oceanbase，来满足其淘宝海量业务对互动式和实时性数据分析与处理的需求。

国内学术界近年来学术水平不断提高，与国外学术交流合作深入开展，所以在各个领域与技术方向都能够很快地掌握并了解国外最新研发的动态和资料。具体到互动式实时性数据分析技术领域，国内的数据库专委会在这方面做了大量的工作，国内数据库的传统优势研发单位，例如清华大学、人民大学、东北大学、哈尔滨工业大学、华东师范大学、西北工业大学、武汉大学、北京航空航天大学、上海交通大学、复旦大学、浙江大学、北京大学的数据库组等都积累了很强的人才储备，并自主研发了很多具有原创性的工作。科技部 973 项目也于 2014 年组织了大数据专项资金，在城市大数据、多媒体大数据、互联网大数据等方向都有专项的资助。各个课题里面所面临的一个共同挑战就是如何快速高效地支持互动式数据分析，并提供实时性的数据处理结果。这方面的研发力量已经在国际高水平的学术会议上发表了很多相关论文，并推动了国内在这个领域相关技术与系统的快速增长。其中中国人民大学 WAMDM 实验室开发的 COLA (Cloud-Based System for Online Aggregation)^[136] 是一个非常代表性的在线分析系统。该系统主要利用在线聚集技术，实现了海量数据上的在线分析。COLA 具有如下的特点：

- COLA 不仅能够支持单表在线聚集，而且支持基于多表连接的在线聚集。
- 系统具有向后兼容性和用户透明性，用户可以直接提交 SQL 查询语句，也可以直接提交批处理方式的 MapReduce 源码，而不需要了解在线查询处理过程和结果估计过程。
- 设计了直观的图形界面展示最新估计结果以及历史结果，并为用户提供了友好的参数设置界面，方便与系统进行交互。
- 在提供聚集结果和置信区间的同时，还提供了完成查询的剩余时间，以及现在停掉查询所能节省的成本，辅助用户做出停止查询与否的决策。

COLA 的系统结构主要包括四个层次：用户接口层、查询处理层、在线聚集处理层以及数据管理层。COLA 既支持用户提交 SQL 查询语句，也支持用户提交原始批处理形式的 MapReduce 脚本，用户提交的请求通过用户接口层传给查询处理层。查询处理层对得到的查询语句进行语法解析，并根据直方图统计信息为查询选择合适的 MapReduce 执行计划，如果用户选择的是在线聚集查询，那么该层还要对 MapReduce 代码进行在线化处理。处理后的 MapReduce 执行计划交给在线聚集处理层，该层按照执行计划将 MapReduce 作业提交给 Hadoop 平台执行，并进行在线聚集的相关计算，包括任务和查询的进程估计、聚集结果估计以及置信区间的计算等。数据管理层主要负责数据的存储以及元数据的管理，同时还负责存储层的数据采样以及直方图数据的收集。

我国在这个方向上的开源系统的开发与推进上与国外相比还有一定差距，尤其是在开源社区的建立、维护和发展上，需要努力将学术成果推广为具有广泛影响力的开源系统与社区。

5.5 国内外研究进展比较

总的来说，国内企业与高校在互动式与实时性数据分析方向以及系统开发方向已经积累了很多经验与人才，也已经做出了很多相关工作和成果。但是与国外现有技术与系统相比，还未能做到大规模开源推广并建立自己的相关社区。国内这方面的研发力量下一步的发展必然是进一步结合国外先进的开源项目，并整合到自主的相关系统中去，从而进一步推动自主研发的技术发展和积累。

5.6 总结

大数据存储、处理和分析对系统的可用性、可拓展性、容错性等一系列关键指标都提出了新的挑战。互动式和实时性的数据处理与分析是满足高效率从大数据中发掘有用信息的一项关键技术。国内外的学术界和工业界在这个方向上都投入了非常大的人力物力，在接下来的几年里这个方向会有爆发式的增长，这其中有几个关键技术有待进一步研究开发。

现有的在线实时性数据分析针对的是基本的单表聚合类操作符，比如 SUM、COUNT、QUANTILE 等，更复杂的分析类型如聚类、矩阵分解、画像等还没有很好的相关解决方案。

在互动式数据分析方向，面向单表的互动式处理已经非常成熟，即使是在数据量很大的情况下，通过索引、采样、分布式并行处理等一系列相关技术，支持单表的互动式分析总的来说已经不是一个技术难题。但对于多表的联合或者关联性很强的操作，比如 join、数据清洗与整合，复杂一些的多数据源数据挖掘与机器学习操作等，都还没有能够完全做到互动式数据分析所要求的高效与快速反应，以及不间断式地返回分析结果的要求。

大数据应用除了数据量大之外，另外两个关键特点是：高维度和多数据源异构数据。在高维度数据上的互动与实时分析技术还有很大欠缺，主要原因是随着数据维度的增长，很多在低维度行之有效数据处理技术都变得与读取并处理全部数据没有太大差别，大大降低了这些技术的有效性。高维度数据的也带来了不同维度之间数据特征提取与关联分析的挑战。多数据源异构数据是对现有技术的另外一大挑战。大数据常常是由不同的应用产生的，这些数据由于来源不同经常包括了结构化数据与非结构化数据，造成了多数据源异构数据的特点。这对互动式实时数据分析技术与系统带来了很大的挑战，如何快速高效地整合清理数据并在多数据源上联合分析发现有用信息是一个根本性的技术难题。

互动式实时性数据分析的主要解决方案集中在分析引擎上，也就是说与传统的数据仓库和 OLAP 应用的出发点很相似。这也带来了一定的局限性，就是要求数据在互动式实时分析过程中要保持静态，不能有数据更新。在这类应用场景和技术框架下支持数据更新，尤其是频繁的数据更新，是一个很大的挑战。例如传统的数据立方体技术在数据

更新之后就有很大的系统维护成本和开销。如何在频繁的数据更新之下还能够高效地支持互动式实时性数据分析是一个需要解决的关键问题，也就是说如何在 OLTP 的环境下，而不仅仅是在 OLAP 的环境下，提供互动式实时性的数据分析解决方案。

6 商业数据管理系统

6.1 引言

数据库技术已经开始逐步摆脱之前关系型数据模型，朝着更为多元化的发展。分布式关系型数据库，NoSQL、NewSQL 都已经占有各自的市场，以此为基础的云端服务^[137]又在弱化其中的差异。

6.2 传统关系型数据库

传统关系型数据库市场依然由 Oracle、MySQL、Microsoft SQL Server 所把控，其目前依然占据数据库市场的最大份额（数据由 DB-Engines Ranking 提供）。然而这三个数据库在产品功能趋同的情况下，也发展差异化。

1. Oracle 数据库

Oracle 是功能最为完善与强大的数据库，可以提供一整套从软件到硬件的各种解决方案。目前，其依然是传统金融、电信行业的重要数据库选型参考。近年来，Oracle 数据库已经不满足提供单纯的数据库软件，开始提供一体机解决方案，这应看成是 Oracle 数据库未来着力发展的一个方向。Oracle 数据库一体机是全面集成了 Oracle 数据库软件和服务器、存储、网络系统的一体化数据库设备。全部作为一个完整的系统设计在一起，无需组装或布线。要使用 Oracle 数据库机，只需打开包装，插上电源线，插上网线，为其命名，然后安装 Oracle 设备管理器软件，即可快速创建一个集群化、高度可用的数据系统。未来，Oracle 一体机数据库主要面向的是中小企业和部门级应用。

2. MySQL 数据库

MySQL 是最为流行的开源数据库产品，随着 Oracle 公司收购 SUN 公司，目前 MySQL 已经隶属于 Oracle 公司。MySQL 是互联网行业使用最为广泛的数据库，Facebook、Google、百度、腾讯、阿里和网易等互联网公司都是其客户。

随着 Oracle 公司在 MySQL 数据库投入的不断增加，MySQL 数据库的功能性在最近几年已经得到极大的提升，并发性与性能也提升巨大，逐步摆脱之前的功能局限，在广泛的领域得到使用。由于 MySQL 是开源的数据库，可以做到真正的自主可控的要求，其目前正在传统行业逐步替换之前的 Oracle 数据库。而随着互联网 + 概念的逐步深入和实际应用，未来将有更多的产品使用 MySQL 数据库。

3. Microsoft SQL Server

Microsoft SQL Server 是一个全面的数据库平台，使用集成的商业智能（BI）工具提供了企业级的数据管理。Microsoft SQL Server 数据库引擎为关系型数据和结构化数据提供了安全可靠的存储功能，可以构建和管理用于业务的高可用和高性能的数据应用程序。Microsoft SQL Server 的优点是可以集成 Windows 平台的所有特性，提供一站式的整体解决方案。缺点为由于是 Windows 数据库，因此其只能部署在 Windows 操作系统上，对于系统的稳定性有所欠缺。同时，也导致其在互联网应用中所占份额相对较少。

4. 总结与对比

下表对比了上述关系型数据库。

	Oracle	MySQL	Microsoft SQL Server
整体功能	非常完善	逐步完善中	完善
发展方向	一体机解决方案	逐步替换传统商业数据库	完整的 Windows 解决方案
产品缺点	商业数据库，价格可能难以承受	需要市场和时间的考验，以此证明是否可以替换传统商业数据库	商业数据库，价格可能难以承受，Windows 平台，稳定性一般
应用范围	传统企业客户	互联网	传统企业客户

6.3 NoSQL

NoSQL(Not Only SQL) 并没有一个准确的定义，但一般认为 NoSQL 数据库应当具有以下的特征：模式自由（schema-free）、支持简易备份（easy replication support）、简单的应用程序接口（simple API）、最终一致性（或者说支持 BASE 特性，不支持 ACID）、支持海量数据（Huge amount of data）等。主流系统包括 MongoDB、Hbase、Cassandra、Redis 等，当然此类系统还十分庞杂，并且不断涌现新的系统。

1. MongoDB

MongoDB 已经超越部分传统关系型数据库，例如 PostgreSQL、DB2，是目前在 IT 行业非常流行的一种非关系型数据库（NoSQL），其灵活的数据存储方式备受当前 IT 从业人员的青睐。

MongoDB 很好地实现了面向对象思想，在 MongoDB 中每一条记录都是一个文档对象。MongoDB 最大的优势在于所有的数据持久操作都无需开发人员手动编写 SQL 语句，直接调用方法就可以轻松地实现 CRUD 操作。

MongoDB 2.8 的最新版本已经实现了对 WiredTiger 存储引擎的支持，提供了文档级别的锁，从而使得性能、压缩性、可用性都得到了极大的提升。

2. HBase

HBase 是一个分布式的、面向列的开源数据库，该技术来源于 Fay Chang 所撰写的 Google 论文“Bigtable：一个结构化数据的分布式存储系统”。就像 Bigtable^[142] 利用了 Google 文件系统（File System）所提供的分布式数据存储一样，HBase 在 Hadoop 之上提供了类似于 Bigtable 的能力。HBase 是 Apache 的 Hadoop 项目的子项目。

HBase 不同于一般的关系数据库，它是一个适合于非结构化数据存储的数据库。另一个不同的是 HBase 是基于列的而不是基于行的模式。

3. Cassandra

Cassandra 是一套开源分布式 NoSQL 数据库系统。它最初由 Facebook 开发，用于储存收件箱等简单格式数据，集 GoogleBigTable 的数据模型与 Amazon Dynamo 的完全分布式的架构于一身。2008 年，Facebook 将 Cassandra 开源。此后，由于 Cassandra 良好的可扩放性，被 Digg、Twitter 等知名 Web 2.0 网站所采纳，成为了一种流行的分布式结构化数据存储方案。

Cassandra 是一个混合型的非关系的数据库，其是介于关系数据库和非关系数据库之间的开源产品，是非关系数据库当中功能最丰富、最像关系数据库的。数据存储采用 bjson 格式，因此可以存储比较复杂的数据类型。

4. Redis

Redis 是一个高性能的 key-value 数据库。Redis 的出现，很大程度补偿了之前 memcached 这类 key/value 存储的不足，在部分场合可以对关系数据库起到很好的补充作用。目前它已经取代传统的 memcached 缓存系统，成为最为流行的 key-value 缓存系统。

相比 Memcached 系统，Redis 也是一个 key-value 系统，但是其支持存储的 value 类型相对更多，包括 string(字符串)、list(链表)、set(集合)、zset(sorted set, 有序集合) 和 hash (哈希类型)。这些数据类型都支持 push/pop、add/remove 及取交集并集和差集及更丰富的操作，而且这些操作都是原子性的。

Redis 支持主从同步。数据可以从主服务器向任意数量的从服务器上同步，从服务器可以是关联其他从服务器的主服务器。这使得 Redis 可执行单层树复制。同步功能对读取操作的可扩展性和数据冗余很有帮助。

5. 总结与对比

有些应用或许并不强烈的依赖关系型数据架构模型，这也是 NoSQL 发展的初衷，NoSQL 相比传统数据库性能更好、可扩展性更高。对于互联网业务的发展，NoSQL 可能是未来的一个主要发展方向。

下表总结对比了当前较流行的 NoSQL 数据库的功能与特性。

	MongoDB	HBase	Cassandra	Redis
数据模型	Document	Column-oriented	Column-oriented	Key- Value
数据存储	普通存储	HDFS	普通存储	内存
Map and Reduce 支持	支持	支持	支持	不支持
压缩	支持	支持	支持	不支持
数据一致性	最终一致性保证	一致性保证	最终一致性保证	无一致性保证
事务支持	不支持	不支持	不支持	不支持
二级索引	支持	支持	支持	不支持
全文索引	不支持	支持	不支持	不支持
地理索引	支持	支持	不支持	不支持

6.4 NewSQL

目前有些比较激进的观点认为“关系数据库已死”，我们认为关系数据库和 NoSQL 并不是矛盾的对立体，而是可以相互补充的、适用于不同应用场景的技术。例如实际的互联网系统往往都是 ACID 和 BASE 两种系统的结合。近些年来，以 Spanner 为代表的若干新型数据库的出现，给数据存储带来了 SQL、NoSQL 之外的新思路。这种融合了一致性和可用性的 NewSQL 或许会是未来大数据存储新的发展方向。典型的系统有 Spanner、OceanBase 等。

1. Spanner

Spanner^[145]是 Google 的全球级的分布式数据库（Globally-Distributed Database）。Spanner 具有高扩展性、多版本（multi-version）、世界级分布（globally-distributed）及同步复制（synchronously-replicated）等特性。

Spanner 立足于高抽象层次，使用 Paxos 协议横跨多个数据集把数据分散到世界上不同数据中心的状态机中。世界范围内响应，出故障时客户副本之间自动切换。当数据总量或服务器的数量发生改变时，为了平衡负载和处理故障，Spanner 自动完成数据的重切片和跨机器（甚至跨数据中心）的数据迁移。

Spanner 可以轻松地横跨数百个数据中心将万亿级数据库行扩展到数百万台机器中。高可靠性更是让应用程序如虎添翼，即使面对大范围的自然灾害，可靠性仍然能得到良好的保障（因为 Spanner 有着世界级的数据转移）。最初的用户来自 F1——使用了美国境内的 5 个拷贝。多数其他应用程序都是在同一个地理区域将数据复制 3~5 份，使用相对独立的故障模式，也就是说多数的应用程序会选择低延迟超过高有效性，只用一两个数据中心来保障数据的可靠性。

目前，谷歌的云服务中还没有提供 Spanner，谷歌正在逐步将部分内部业务迁移到 Spanner 上，例如谷歌广告业务，相信在不久的将来，会看到 Google Cloud 正式推出 Spanner 云服务。

2. OceanBase

OceanBase^[146]是一个支持海量数据的高性能分布式数据库系统，实现了数千亿条记录、数百 TB 数据上的跨行跨表事务，由淘宝核心系统研发部、运维、DBA、广告、应用研发等部门共同完成。在设计和实现上，OceanBase 暂时摒弃了不紧急的 DBMS 的功能，例如临时表、视图（view），研发团队把有限的资源集中到关键点上，当前 OceanBase 主要解决数据更新一致性、高性能的跨表读事务、范围查询、join、数据全量及增量 dump、批量数据导入。

目前 OceanBase 已经应用于淘宝收藏夹，用于存储淘宝用户收藏条目和具体的商品、店铺信息，每天支持 4 000~5 000 万的更新操作。目前 OceanBase 还处于阿里集团内部推广应用的阶段，随着在内部系统上的逐渐稳定，后续阿里云可能会考虑提供 OceanBase 的云服务。

3. NewSQL 与分布式数据库对比

下表显示了国内 NewSQL 的代表 OceanBase 与分布式 MySQL 数据库之间的对比：

	OceanBase	分布式 MySQL
访问接口	API 接口、MySQL 客户端协议	MySQL 客户端协议
数据访问透明	是	是对 SQL 语句有限制
数据一致性	强一致性	最终一致性
高可用	Paxos 协议	通过 Proxy 和数据库复制技术
性能瓶颈	Update Server 是单点瓶颈	无，视数据库单机容量
应用类型	特定应用类型，比如每日更新量较小，可以被 Update Server 完全缓存	通用业务类型

6.5 云关系数据库

随着云计算的快速兴起，数据库作为提供海量结构化和非结构化数据存储、管理和分析的后端系统进入一个崭新的发展阶段。包括谷歌、亚马逊、微软，以及国内的阿里云在内的各大云计算服务厂商纷纷推出了自己的云关系数据库相关的产品。相比传统数据库，云数据库具有低成本、易运维、可伸缩、高可用等优势。

1. Amazon RDS

Amazon RDS (Amazon Relational Database Service)^[140]是亚马逊提供的关系数据库服务，是一个部署在云端的 Web 服务，具有容易部署、操作和扩展等特点。它可以将用户从费时的数据库管理任务中解放出来，提供了具有成本效益和容量动态调整的能力，从而使用户专注于自己的应用程序和业务。

AmazonRDS 支持的数据库类型包括 MySQL、PostgreSQL、Oracle、SQLServer 等几乎所有主流关系型数据。支持跨可用域部署，承诺 SLA 99.95%。硬盘存储能力支持用户自定制，根据用户选择的存储能力收费。支持 ScaleUp、读写分离等数据库扩展模式，暂不支持数据库的横向扩展。

2. Google Cloud SQL

GoogleCloud SQL^[143]是谷歌推出的基于 MySQL 的在线云关系数据库服务。谷歌负责处理 MySQL Replication、版本补丁升级、数据库管理等工作，确保用户的数据库稳定和高性能。

GoogleCloud SQL 完全兼容 MySQL 协议，支持跨物理区域的数据复制、自动备份与故障恢复、读写分离等特性，但限制部分 SQL 的使用和 UDF 等部分功能。

3. Azure SQL Database

Azure SQL Database^[144]是微软推出的一种以服务的方式提供的关系数据库服务，支持 SQL Server、MySQL 两种数据库。Azure SQL Database 的主要特色有：垂直与横向伸缩、提供 99.99% 的可用性 SLA、主动地域复制、自动软件修补等。

4. 阿里云 RDS

阿里云的云数据库是构建在 SSD 盘上，完全兼容 MySQL、SQLServer、PostgreSQL 协议的关系型数据库服务。采取主从双机热备架构，具有多重安全防护措施和完善的性能监控体系，并提供专业的数据库备份、恢复及优化方案，使用户能专注于应用开发和业务发展。

5. 网易 RDS

网易云数据库（Netease RDS）是由网易数据库技术团队基于开源的 MySQL 打造的关系数据库云托管平台，构建于网易私有云 IaaS 服务之上，面向网易众多的互联网和移动终端产品，提供高可用、高可靠、高性能、可扩展的在线数据库服务。

网易云 RDS 主要有以下特点：双机热备、同步复制保证数据完全不丢，FlashBack 和 Point-in-time 恢复，在线垂直伸缩，TopSQL 分析，增量备份。

6. 总结与对比

当前云关系数据库的主要以 MySQL 数据库为主，部分云数据库服务公司还提供有 Oracle、Microsoft SQL Server、PostgreSQL 等数据库支持。

下表总结了当前各主要关系型云数据库的功能对比。

	Amazon RDS	Google CloudSQL	Azure SQL Database	阿里云 RDS	网易云 RDS
只读从节点	支持	支持	支持	支持	支持
数据一致性	支持 通过底层块设备实现	不支持	不支持	不支持	支持 通过对数据库进行额外的开发
数据迁移功能	不支持	不支持	不支持	支持	支持
闪回功能	不支持	不支持	不支持	不支持	支持
数据库支持	MySQL、Oracle、PostgreSQL、Microsoft SQL Server、Aurora	MySQL	Microsoft SQL Server、MySQL	MySQL、PostgreSQL	MySQL
SLA	99. 95%	99. 95%	99. 9%	99. 95%	99. 95%

6.6 NoSQL 的云数据库

很多云计算的服务厂商推出了 NoSQL 的云数据库产品，其中，以 GoogleBigTable 云服务、AmazonSimpleDB 和 Amazon DynamoDB 服务应用最为广泛。

1. Google BigTable

GoogleCloud BigTable 云服务是谷歌基于内部的 BigTable 系统实现的一种以列簇式存储，将同一列数据存在一起，查找速度快，可扩展性强，更容易进行分布式扩展的云服务。它为用户提供了一个快速的、完全自动化管理的、可以无限扩展的 NoSQL 服务，可用于 Web、移动端、互联网应用，处理 TB ~ PB 级数据。已在谷歌内部已使用 BigTable

十余年，包括 Google analytics、Gmail 在内的产品均在其上构建。

2. Amazon DynamoDB

Amazon DynamoDB^[141]是一项快速灵活的 NoSQL 数据库云服务，适合所有需要一致性且延迟低于 10 毫秒的任意规模的应用程序。它是完全托管的云数据库，支持文档和键值存储模型。其灵活的数据模型和可靠的性能令其成为移动、Web、游戏、广告技术、物联网和众多其他应用的不二之选。

3. Amazon SimpleDB

Amazon SimpleDB 是一种可用性高、灵活性大的非关系数据存储，能够减轻数据库管理的工作。开发人员只需通过 Web 服务请求执行数据项的存储和查询，Amazon SimpleDB 将负责余下的工作。

Amazon SimpleDB 最突出的特色是会自动创建和管理分布在多个地理位置的数据副本，以此提高可用性和数据持久性。

4. 总结与对比

有些应用或许并不强烈的依赖关系型数据架构模型，这也是 NoSQL 发展的初衷，NoSQL 相比传统数据库性能更好，可扩展性更高。对于互联网业务的发展，NoSQL 可能是未来的一个主要发展方向。

下表总结了当前 NoSQL 数据库功能与特性之间的对比：

	Google Big Table	Amazon DynamoDB	Amazon SimpleDB
数据模型	Column-oriented	Key-Value	Key-Value
数据存储	GFS2	SSD	普通存储
Map and Reduce 支持	支持	不支持	不支持
压缩	支持	不支持	不支持
数据一致性	是	是	不支持
事务支持	支持	支持	不支持
二级索引	支持	不支持	不支持
全文索引	支持	不支持	不支持
地理索引	支持	不支持	不支持
访问方式	API 接口	API 接口	API 接口、Rest 接口

6.7 总结

传统关系型数据库依然把持着大部分的市场份额，Oracle、MySQL、Microsoft SQL Server 依然使用最为广泛的数据库系统。但是三者的发展侧重点有所不同。Oracle 的发展侧重于软硬件一体机的产品化发展路线，MySQL 现阶段还处于功能完善阶段，使其成为新型系统的数据库首选。Microsoft SQL Server 发展主要是提供基于 Windows 的一站式软件集成解决方案。

传统数据库还是受限于关系模型，使其模式在某些应用中显得僵硬和缓慢。最

近几年出的 NoSQL 数据库得到了广泛应用，其主要特点是 schema free^[139]。分布式是对于传统数据库在数据模型和扩展性方面的一个重要补充。这其中以 MongoDB 最为典型，已经超越传统 DB2 和 PostgreSQL 数据库，成为最为流行的 NoSQL 数据库。

传统数据库对于大数据下 OLAP 查询同样存在缺陷，因此以 Hadoop 为代表的分布式存储与 Map Reduce 模型成为新的大数据查询模型^[138]。Hive 在 Hadoop 基础上提供了 SQL 接口，使得传统开发人员避免了额外的学习成本，应用也能更为平滑地迁移到 Hadoop。而 Spark SQL 优化了 Hive 的架构，使其拥有更好的性能。

随着社交网络的流行，越来越多的关系需要以图的方式进行存储，这使得 Neo4j 图数据库在近些年得到广泛应用。而分布式的图数据库 FlockDB 在可扩展性方面做得更好，未来或许会飞速发展。

NoSQL 虽然解决了当前互联网高并发、大规模访问的性能瓶颈，但是通过牺牲事务一致性、简化查询和数据结构，必然会有业务局限性。随着互联网业务的发展，新的业务需求希望可以保持对事务的支持，同时能够处理海量数据的复杂的多表关联剖析、即席查询等。Google Spanner 和阿里的 OceanBase 为其中的代表，相信在未来会有更多的类似产品诞生。

随着云计算的兴起，众多厂商也推出了云关系数据库或 NoSQL 数据库产品。虽然在 DBaaS 中仍存在许多问题，尤其是关于存储在云的敏感信息，以及服务中断的问题，还有包括向云端数据库迁移困难，没有足够成熟的案例等问题，不过，云数据库和工具这一新兴市场仍在加速发展中，云数据库无疑是未来发展的大势。

7 发展趋势

大数据的出现给传统数据管理技术带来了众多挑战。本报告上述内容正是近年来针对数据管理的最新研究成果，且基于这些成果构建了若干实际可行的新型系统。但随着数据规模以及应用需求的进一步发展，未来的数据管理技术仍旧面临着新的问题和挑战。海量信息从不同的数据源产生，如何将这些数据进行融合并从中提取出有价值的知识会是未来大数据管理的一个重点关注内容。在大数据时代，传统的计算机系统架构已经很难满足处理需求，必须利用新的硬件技术来提升系统效能。从目前的趋势来看，新型处理器以及新型存储设备是主流研究方向。传统的针对小数据的被动式保护方式在大数据环境下已经不再适用，因此隐私管理将是未来大数据管理的一个重要问题。

7.1 数据融合与知识融合

随着计算机技术的深入发展和广泛应用，海量数据和信息从各种应用和数据源中源源不断地产生出来，人类社会开始进入大数据时代。大数据给各行各业以及社会服务带

来潜在的价值。基于大数据的分析可以提高企业的利润、改善社会服务（如医疗、教育、交通等）和方便个人决策，等等。因而，大数据时代的核心问题是如何能全面而深层次地挖掘和利用数据中的价值。大数据除了具有海量的特点外，还具有多源、动态、异构和异质等特点。通常要更好地挖掘出数据中潜在的价值需要将不同数据源甚至不同领域的数据相互关联、融合。这就需要新的面向大数据有效分析和利用的数据融合和知识融合技术。针对大数据分析的典型特点，新的数据融合和知识融合技术将主要面临以下几个方面的挑战。

1) 多源性。大数据不仅数据量大，而且不同数据源的数目也很大。据调查，Web 上同一领域（如餐馆、图书、酒店等）就有十多万个不同数据源。

2) 动态性。许多数据源提供的内容是动态变化的，如股票信息等。这就需要适应动态数据的算法，如联机算法、增量更新算法，等等。

3) 异构性。来自不同数据源的数据通常具有不同的结构或模式信息，或者不同数据源提供的可能分别是无结构、半结构或结构化数据。

4) 多模态。不同数据源可能提供不同模态的数据，如有的是文本数据，有的是图像，而有的是音频或者视频，等等。

5) 跨语言。不同数据源可能提供不同语言的数据，如中文的百度百科和英文的维基百科，等等。

6) 异质性。在大数据环境中，不同数据源提供的数据的质量不同。因而在数据和知识融合过程中要对数据的质量和可信度进行检测和评估，并保证数据是可溯源的。

7) 安全性。大数据分析会把不同数据源的信息进行关联和归纳，从而形成新的认知和知识。如果这些内容是用户的个人隐私甚至国家机密，就会造成严重危害。所以数据和知识融合过程必须保证敏感内容及个人隐私等不被泄露。

8) 可适性。大数据的数据融合和知识融合是面向大数据分析的，所以也可以采用一些端到端的技术，如机器学习技术，在融合的过程中直接产生分析结果。另外，数据分析的结果也可以反过来影响数据融合和知识融合的过程。融合的过程应该可以根据实际数据分析的需求做调整，并根据数据分析的结果做进一步改进从而能得到更好的分析结果。

7.2 基于新型硬件的大数据管理

随着社交网、物联网等技术的快速发展，数据的产生量呈现爆炸性的增长。面对如此巨大的数据量，对数据的有效管理成为一个很重要的问题。当前的大数据管理系统是基于传统的 CPU-内存-硬盘的体系结构，在这种体系结构中，CPU 和磁盘的速度严重不匹配，在过去的几十年，CPU 的处理速度提升了近千倍，而磁盘的速度则没有获得很大的提升。同时，当前的大数据处理系统也都是基于单核 CPU 的。因此在大数据管理中很有必要引入新型硬件。然而，引入新型硬件之后会带来如下的挑战。

1) 基于多核 CPU 的大数据管理。单核 CPU 的发展一直遵循着摩尔定律，但由于功

耗、互联线延时和设计复杂度的物理条件限制，其架构已经从单核转入多核。如果说传统的运行在单核 CPU 上的应用程序，可以依赖摩尔定律的规律性直接获得来自硬件的性能提升；在现今各大处理器厂商推出的多核处理器上，应用程序需要优化代码从而充分利用多个处理器核的并行计算能力。相应地，大数据管理程序也要优化代码从而充分发挥多核 CPU 的处理能力。

2) 基于新型存储介质的大数据管理。新型存储介质如 SSD、PCM 其访问速度比磁盘快很多，同时又具有非易失性，引入这些存储介质能加快数据访问速度。引入这些存储介质后，未来的存储系统中可能会出现 DRAM、PCM、SSD 和 HDD 等不同存储介质共存的局面。但是不同的存储介质在易失性、访问延迟、存取模式、价格等方面存在很大的差别，如何发挥新型存储器件在访问性能、存取模式等方面的优势，构建基于新型存储介质的大数据存储是大数据管理的首要问题。同时也需要优化大数据的查询和处理等算法从而最大限度的利用新型存储介质的特性。

3) 低能耗问题。在当今的节能减耗的时代背景下，如何降低数据中心的能耗也是一个重要的问题。在大数据管理系统中，能耗主要由两大部分组成：硬件能耗和软件能耗，二者之中又以硬件能耗为主。SSD 和 PCM 都具有低能耗的优点，因此如何利用新型硬件降低系统的能耗也是一个挑战性的问题。

7.3 大数据隐私管理

大数据的处理流程涉及数据收集、数据集成与融合、数据分析，以及数据解释四个步骤。其中数据收集包括公开数据和私有数据的收集；数据集成与融合主要处理数据之间的冗余、不一致、相互拷贝关系等问题；数据分析的目的是从数字化数据与模拟化数据中抽取或者学习有价值的模型和规则；而数据解释主要是通过可视化、数据溯源等技术来展示大数据的分析结果。然而，在大数据的整个处理框架和生命周期中，每个步骤均存在披露和破坏数据隐私的风险。因此解决大数据隐私问题的当务之急是，针对不同的风险，建立混合式与综合性的隐私管理框架，并积极拓展隐私管理的关键技术研究。大数据带来的隐私风险主要有以下四点。

1) 数据肆意收集带来的风险。在大数据环境中，可以通过医疗就医记录、购物及服务记录、网站搜索记录、手机通话记录、手机位置轨迹记录等来获取用户的信息。如果个人数据被不可信的第三方服务收集，则个人隐私很有可能被泄露或者卖给恶意攻击者。例如，不可信的位置服务恶意收集用户的位置信息，则用户的敏感位置可能会被披露。

2) 数据集成融合带来的风险。集成融合通常采用链接操作使多个异构数据源汇聚在一起，并且识别出相应的实体。小数据源通常能够反映出用户的某个活动，比如接受的医疗、购买的商品、搜索的网站、手机留下的位置特征、与社交网络互动信息、政治活动等。而多个数据源的集成和融合几乎能过推理出个人所有的敏感信息，无形中给个人隐私的保护带来严峻挑战。

3) 数据分析带来的风险。大数据环境下，以 Hadoop + MapReduce、Storm、Dremel，

以及 R + Hadoop 为代表的强大计算框架，能够以批处理或者流式处理的方式并行处理大规模数据；高性能算法不但能够深层分析大数据中那些细小的、彼此之间毫无关联的数据碎片，同时也为恶意分析者提供了确凿的攻击背景知识，使他们能够通过分析挖掘大数据中的隐私信息。

4) 数据解释带来的风险。在数据解释步骤中可能存在前景知识攻击、通过数据溯源图挖掘元数据之间的依赖关系等。

8 结束语

大数据给整个社会带来了巨大的变革，也给相关的技术领域带来了巨大的挑战。不同领域的学者均尝试从自身的角度出发来解决大数据的种种问题。数据库作为数据的重要载体，比其他领域更加直面大数据的挑战。关系数据库的提出和发展，在过去几十年里有效地推动了其他学科的进步。现有的数据库方法常常更关注处理方法的效率 (efficiency)，比如通过各种优化技术来加速查询处理。但大数据所面对的不仅仅是简单的 SPJ 查询，因此传统的数据库方法如何去适应并有效地处理大数据是值得深思的问题。大数据时代我们首要关注的应当是“功效” (efficacy) 而不是效率，因为对于大数据“失效”的数据库根本就无从谈效率问题。希望本报告的内容能对同行有一些有益的启发和思考，并在未来真正推动大数据时代的数据管理研究。

致谢

衷心感谢参与本报告撰写、编辑与审校的人员！

参与本报告编写的人员包括（按章节排序）：孟小峰教授（中国人民大学，概述、发展趋势）、高军教授与崔斌研究员（北京大学，图数据管理系统）、禹晓辉教授（山东大学，流数据管理系统）、李国良副教授（清华大学，众包数据管理系统）、李飞飞副教授（美国犹他大学，在线数据分析与管理系统）、陈刚教授（浙江大学，商业数据管理系统）、王秋月博士（中国人民大学，发展趋势）。

中国人民大学孟小峰教授统稿并审校了全文。博士生王春凯、慈祥参与了本报告的排版和编辑工作。

参考文献

图数据管理系统

[1] Aapo Kyrola, Guy Blelloch, and Carlos Guestrin. GraphChi: large- scale graph computation on just a PC

- [C]. Proceedings of the 10th USENIX conference on Operating Systems Design and Implementation , OSDI , 2012 , 31-46.
- [2] Alekh Jindal , Praynaa Rawlani , Eugene Wu , Samuel Madden , Amol Deshpande , Mike Stonebraker . VERTEXICA : Your Relational Friend for Graph Analytics ! [J]. PVLDB 7(13) , 2014 , 1669-1672 .
- [3] Amitabha Roy , Ivo Mihailovic , and Willy Zwaenepoel . X-stream : Edge- centric graph processing using streaming partitions [J]. In Proceedings of the Twenty- Fourth ACM Symposium on Operating Systems Principles . SOSP , 2013 , 472-488 .
- [4] Chang Zhou , Jun Gao , Binbin Sun , Jeffrey Xu Yu . MOCgraph : Scalable Distributed Graph Processing Using Message Online Computing [J]. PVLDB 8(4) , 2014 , 377-388 .
- [5] Ching Avery . Giraph : Large- scale graph processing infrastructure on Hadoop [R]. In Proceedings of Hadoop Summit . , Santa Clara , USA ; 2011 .
- [6] Grzegorz Malewicz , Matthew H Austern , Aart JC Bik , James C Dehnert , Ilan Horn , Naty Leiser , Grzegorz Czajkowski . Pregel : a system for large-scale graph processing [C]. Proceedings of the 2010 ACM SIGMOD International Conference on Management of data . ACM , 2010 , 135-146 .
- [7] Isabelle Stanton , Gabriel Kliot . Streaming graph partitioning for large distributed graphs [C]. In Proceedings of the 18th ACM SIGKDD International Conference .
- [8] Jeffrey Dean , Sanjay Ghemawat . MapReduce : simplified data processing on large clusters [C]. Communications of the ACM , 2008 , 51(1) : 107-113 .
- [9] Jiefeng Cheng , Qin Liu , Zhenguo Li , Wei Fan , John C. S. Lui , Cheng He . VENUS : Vertex- centric streamlined graph computation on a single PC [R]. ICDE , 2015 : 1131-1142 .
- [10] Jiwon Seo , Stephen Guo , Monica S. Lam . SociaLite : Datalog extensions for efficient social network analysis [R]. ICDE , 2013 : 278-289 .
- [11] Joseph E Gonzalez , Yucheng Low , Haijie Gu , Danny Bickson , Carlos Guestrin . Powergraph : distributed graph- parallel computation on natural graphs [J]. In Proc. of OSDI , 2012 ; 17-30 .
- [12] Jun Gao , Jiashuai Zhou , Chang Zhou , Jeffrey Xu Yu . GLog : A high level graph analysis system using MapReduce [R]. ICDE , 2014 : 544-555 .
- [13] JunZhang , Chaokun Wang , Yuanchi Ning , Yichi Liu , Jianmin Wang , Philip Yu . LaFT- Explorer : Inferring , Visualizing and Predicting How Your Social Network Expands (system demo) [M]. Chicago : ACM Press , 2013 .
- [14] Lei Zou , Jinghui Mo , Lei Chen , M. Tamer Özsu , Dongyan Zhao . gStore : Answering SPARQL Queries Via Subgraph Matching [J]. Proc. VLDB 4(8) , 2011 : 482-493 .
- [15] Lei Zou , Ruizhe Huang , Haixun Wang , Jeffrey Xu Yu , Wenqiang He , Dongyan Zhao . Natural language question answering over RDF : a graph data driven approach [C]. SIGMOD Conference , 2014 : 313-324 .
- [16] QunChen , Song Bai , Zhanhuai Li , Zhiying Gou , BoSuo , Wei Pan . GraphHP : A Hybrid Platform for Iterative Graph Processing [OL] . [2014] . <http://wowbigdata.cn/C-paper.htm> .
- [17] Semih Salihoglu , Widom Jennifer . Optimizing graph algorithms on pregel- like systems [C]. Proceedings of the VLDB Endowment . 2014 , 7(7) : 577-588 .
- [18] Semih Salihoglu , Widom Jennifer . Optimizing graph algorithms on pregel- like systems [C]. Proceedings of the VLDB Endowment . 2014 , 7(7) : 577-588 .
- [19] Semih Salihoglu , Jennifer Widom . GPS : a graph processing system [C]. SSDBM , 2013 : 22 ; 1-22 ; 12 .

- [20] Ye Yuan, Guoren Wang, Jeffery Yu Xu, Lei Chen. Efficient distributed subgraph similarity matching [J]. VLDB J, 2015, 24(3) : 369-394.
- [21] Yingxia Shao, Bin Cui, Lei Chen, Lin Ma, Junjie Yao, Ning Xu. Parallel subgraph listing in a large-scale graph[C]. SIGMOD Conference, 2014: 625-636.
- [22] Yuanyuan Tian, Andrey Balmin, Severin AndreasCorsten, Shirish Tatikonda, John McPherson. From "think like a vertex" to "think like a graph" [J]. Proceedings of the VLDB Endowment, 2013 , 7 (3) :193-204.
- [23] Yucheng Low, Danny Bickson, Joseph Gonzalez, Carlos Guestrin, Aapo Kyrola, Joseph M. Hellerstein. Distributed graphlab: A framework for machine learning and data mining in the cloud[J]. Proceedings of the VLDB Endowment, 2012 , 5(8) :716-727.
- [24] <http://www.cdblp.cn/>.
- [25] <http://www.scholat.com/>.
- [26] Robert Ryan McCune, Tim Weninger, Greg Madey. Thinking Like a Vertex: a Survey of Vertex-Centric Frameworks for Large- Scale Distributed Graph Processing ACM Computing Surveys [OL]. http://www3.nd.edu/~rmccune/papers/Think_Like_a_Vertex_MWM.pdf.
- [27] Joseph E. Gonzalez, Reynold S. Xin, Ankur Dave, Daniel Crankshaw, Michael J. Franklin, Ion Stoica. GraphX: Graph Processing in a Distributed Dataflow Framework[J]. OSDI, 2014: 599-613.

流数据管理系统

- [28] Ulf Schreier, Hamid Pirahesh, Rakesh Agrawal, C. Mohan. Alert. An architecture for transforming a passive DBMS into an active DBMS[C]. In 17th International Conference on Very Large Data Bases, Barcelona, Catalonia, Spain:1991 : 469-478.
- [29] Nicholas J Belkin, W Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? [J]. Commun. ACM , 1992,35(12) :29-38.
- [30] Ben Kao, Hector Garcia- Molina. An overview of real- time database systems[J]. Real Time Computing, 1994:261-282.
- [31] Norman W Paton, Oscar Dáz. Active database systems[J]. ACM Comput. Surv. 1999 , 31(1) :63-103.
- [32] Daniel J. Abadi, Donald Carney, Ugur Çetintemel, Mitch Cherniack, Christian Convey, Sangdon Lee, Michael Stonebraker, Nesime Tatbul, Stanley B. Zdonik. Aurora: a new model and architecture for data stream management[J]. VLDB J. , 2003,12(2) :120-139.
- [33] Sirish Chandrasekaran, Owen Cooper, Amol Deshpande, Michael J. Franklin, Joseph M. Hellerstein, Wei Hong, Sailesh Krishnamurthy, Samuel Madden, Vijayshankar Raman, Frederick Reiss, Mehul A. Shah. Telegraphcq: Continuous dataflow processing for an uncertain world[J]. CIDR , 2003.
- [34] Arvind Arasu, Brian Babcock, Shivnath Babu, John Cieslewicz, Mayur Datar, Keith Ito, Rajeev Motwani, Utkarsh Srivastava, Jennifer Widom. Stream: The stanford data stream management system[J]. Book chapter,2004.
- [35] Mitch Cherniack, Hari Balakrishnan, Magdalena Balazinska, Donald Carney, Ugur Çetintemel, Ying Xing, Stanley B Zdonik. Scalable distributed stream processing[J]. CIDR , 2003.
- [36] Mehul A. Shah, Joseph M. Hellerstein, Eric A. Brewer. Highly- available, fault- tolerant, parallel dataflows[C]. Proceedings of the ACM SIGMOD International Conference on Management of Data, Paris, France:2004 :827-838.

- [37] Daniel J Abadi, Yanif Ahmad, Magdalena Balazinska, Ugur Çetintemel, Mitch Cherniack, Jeong-Hyon Hwang, Wolfgang Lindner, Anurag Maskey, Alex Rasin, Esther Ryvkina, Nesime Tatbul, Ying Xing, Stanley B. Zdonik. The design of the borealis stream processing engine[J]. CIDR, 2005: 277-289.
- [38] <http://hadoop.apache.org/>.
- [39] Leonardo Neumeyer, Bruce Robbins, Anish Nair, Anand Kesari. S4: Distributed stream computing platform[J]. ICDM Workshops, 2010:170-177.
- [40] Ankit Toshniwal, Siddarth Taneja, Amit Shukla, Karthikeyan Ramasamy, Jignesh M. Patel, Sanjeev Kulkarni, Jason Jackson, Krishna Gade, Maosong Fu, Jake Donham, Nikunj Bhagat, Sailesh Mittal, Dmitriy V. Ryaboy. Storm@twitter[C]. SIGMOD Conference,2014:147-156.
- [41] Matei Zaharia, Tathagata Das, Haoyuan Li, Scott Shenker, Ion Stoica. Discretized streams: An efficient and fault-tolerant model for stream processing on large clusters[C]. 4th USENIX Workshop on Hot Topics in Cloud Computing, HotCloud'12, Boston, MA, USA;2012:12-13.
- [42] <http://samza.incubator.apache.org/>.
- [43] Tyler Akidau, Alex Balikov, Kaya Bekiroglu, Slava Chernyak, Josh Haberman, Reuven Lax, Sam McVeety, Daniel Mills, Paul Nordstrom, Sam Whittle. Millwheel: Fault- tolerant stream processing at internet scale[J]. PVLDB, 2013,6(11):1033-1044.
- [44] Kulkarni S, Bhagat N, Fu M, et al. Twitter Heron: Stream Processing at Scale[C]. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, 2015: 239-250.
- [45] William Douglas Clinger. Foundations of actor semantics. 1981.
- [46] Patrick Hunt, Mahadev Konar, Flavio Paiva Junqueira, Benjamin Reed. Zookeeper: Wait- free coordination for internet-scale systems[C]. In USENIX Annual Technical Conference, 2010.
- [47] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: cluster computing with working sets [C]. Proceedings of the 2nd USENIX conference on Hot topics in cloud computing,2010.
- [48] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing [C]. In Proceedings of the 9th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2012, San Jose, CA, USA , 2012:15-28.
- [49] Jay Kreps, Neha Narkhede, Jun Rao, et al. Kafka: A distributed messaging system for log processing [C]. In Proceedings of the NetDB, 2011.
- [50] <http://hadoop.apache.org/docs/r2.2.0/hadoop-yarn/>.
- [51] <http://leveldb.org/>.
- [52] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Michael Burrows, Tushar Chandra, Andrew Fikes, Robert Gruber[C]. Bigtable: A distributed storage system for structured data (awarded best paper!) [C]. In 7th Symposium on Operating Systems Design and Implementation (OSDI '06), Seattle, WA, USA , 2006:205-218.
- [53] James C Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J J Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson C 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 [J]. ACM Trans. Comput. Syst., 2013, 31(3):8.
- [54] <http://mesos.apache.org/>.
- [55] <http://zookeeper.apache.org/>.
- [56] Rajagopal Ananthanarayanan, Venkatesh Basker, Sumit Das, Ashish Gupta, Haifeng Jiang, Tianhao Qiu, Alexey Reznichenko, Deomid Ryabkov, Manpreet Singh, and Shivakumar Venkataraman. Photon: fault-tolerant and scalable joining of continuous data streams [C]. In SIGMOD Conference, 2013: 577-588.
- [57] Diego Ongaro, Stephen M. Rumble, Ryan Stutsman, John K. Ousterhout, Mendel Rosenblum. Fast crash recovery in ramcloud [C]. In Proceedings of the 23rd ACM Symposium on Operating Systems Principles 2011, SOSP 2011, Cascais, Portugal, October 23-26, 2011, 2011:29-41.
- [58] Hyeontaek Lim, Bin Fan, David G. Andersen, Michael Kaminsky. SILT: a memory-efficient, high-performance key-value store [C]. In Proceedings of the 23rd ACM Symposium on Operating Systems Principles 2011, SOSP 2011, Cascais, Portugal, 2011:1-13.
- [59] <http://tachyon-project.org/>.
- [60] Miyuru Dayarathna, Toyotaro Suzumura. Hirundo: a mechanism for automated production of optimized data stream graphs [C]. In Third Joint WOSP/SIPEW International Conference on Performance Engineering, ICPE'12, Boston, MA, USA, 2012:335-346.
- [61] Mehul A Shah, Joseph M Hellerstein, Sirish Chandrasekaran, Michael J. Franklin. Flux: An adaptive partitioning operator for continuous query systems [C]. In Proceedings of the 19th International Conference on Data Engineering, Bangalore, India, 2003:25-36.
- [62] Goetz Graefe. Encapsulation of parallelism in the volcano query processing system [C]. In Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, 1990: 102-111.
- [63] Mehul A Shah, Joseph M Hellerstein, Sirish Chandrasekaran, Michael J. Franklin. Flux: An adaptive partitioning operator for continuous query systems [C]. In Proceedings of the 19th International Conference on Data Engineering, Bangalore, India, 2003:25-36.
- [64] Leonardo Aniello, Roberto Baldoni, Leonardo Querzoni. Adaptive online scheduling in storm [C]. In The 7th ACM International Conference on Distributed Event-Based Systems, DEBS '13, Arlington, TX, USA, 2013:207-218.
- [65] Matthias J Sax, Malú Castellanos, Qiming Chen, Meichun Hsu. Aeolus: An optimizer for distributed intra-node-parallel streaming systems [J]. In ICDE, 2013:1280-1283.
- [66] Nasir M A U, Morales G D F, García-Soriano D, et al. The Power of Both Choices: Practical Load Balancing for Distributed Stream Processing Engines [J]. arXiv preprint arXiv:1504.00788, 2015.
- [67] Jeong-Hyon Hwang, Magdalena Balazinska, Alex Rasin, Ugur Çetintemel, Michael Stonebraker, Stanley B. Zdonik. High-availability algorithms for distributed stream processing [J]. In ICDE, 2015: 779-790.
- [68] Jeong-Hyon Hwang, Ugur Çetintemel, Stanley B. Zdonik. Fast and highly-available stream processing over wide area networks [C]. In Proceedings of the 24th International Conference on Data Engineering, Cancún, México, 2008:804-813.
- [69] Magdalena Balazinska, Hari Balakrishnan, Samuel Madden, Michael Stonebraker. Fault-tolerance in the borealis distributed stream processing system [C]. In Proceedings of the ACM SIGMOD International

- Conference on Management of Data, Baltimore, Maryland, USA, 2005:13-24.
- [70] Gradvohl A L S, Senger H, Arantes L, et al. Comparing Distributed Online Stream Processing Systems Considering Fault Tolerance Issues [J]. Journal of Emerging Technologies in Web Intelligence, 2014, 6 (2) : 174-179.
- [71] Liu Q, Lui J, He C, et al. SAND: A Fault-Tolerant Streaming Architecture for Network Traffic Analytics [C]. Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on. IEEE, 2014: 80-87.
- [72] Castro Fernandez R, Migliavacca M, Kalyvianaki E, et al. Integrating scale out and fault tolerance in stream processing using operator state management [C]. Proceedings of the 2013 ACM SIGMOD international conference on Management of data. ACM, 2013: 725-736.
- [73] <http://data.epfl.ch/squall>.
- [74] Bifet A, Frank E. Sentiment knowledge discovery in twitter streaming data [C]. Discovery Science. Springer Berlin Heidelberg, 2010: 1-15.
- [75] Bifet A, Holmes G, Pfahringer B, et al. Detecting Sentiment Change in Twitter Streaming Data [C]. WAPA. 2011: 5-11.
- [76] Kim H G, Lee S, Kyeong S. Discovering hot topics using Twitter streaming data social topic detection and geographic clustering [C]. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE, ACM International Conference on. IEEE, 2013: 1215-1220.
- [77] Wang H, Can D, Kazemzadeh A, et al. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle [C]. Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012: 115-120.
- [78] Cao J, Zhou Y, Wu M. Adaptive Grid-Based k- median Clustering of Streaming Data with Accuracy Guarantee [C]. Database Systems for Advanced Applications. Springer International Publishing, 2015: 75-91.
- [79] Yanai K, Takamu T, Kawano Y. Real-time Photo Mining from the Twitter Stream: Event Photo Discovery and Food Photo Detection [C]. Multimedia (ISM), 2014 IEEE International Symposium on. IEEE, 2014: 295-302.
- [80] Liwei Lin, Xiaohui Yu, Nick Koudas. Pollux: towards scalable distributed real-time search on microblogs [C]. In EDBT, 2013:335-346.
- [81] 亓开元, 赵卓峰, 房俊, 等. 针对高速数据流的大规模数据实时处理方法 [J]. 计算机学报, 2012, 35(3) : 477-490.
- [82] Z Yu, Y Liu, X Yu, K Pu. Scalable distributed processing of k nearest neighbor queries over moving objects [J]. TKDE, 2014.
- [83] Xiong C, Zhang P, Li Y, et al. A Memory-Based Continuous Query Index for Stream Processing [C]. Big Data (BigData Congress), 2014 IEEE International Congress on. IEEE, 2014: 768-769.
- [84] Duan Q, Wang P, Wu M X, et al. Approximate query on historical stream data [C]. Database and Expert Systems Applications. Springer Berlin Heidelberg, 2011: 128-135.
- [85] Deng Z, Wu X, Wang L, et al. Parallel processing of dynamic continuous queries over streaming data flows [J]. Parallel and Distributed Systems, 2015, 26(3) : 834-846.
- [86] Brian Babcock, Shivnath Babu, Mayur Datar, Rajeev Motwani, Jennifer Widom. Models and issues in data stream systems [C]. In Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on

Principles of Database Systems, Madison, Wisconsin, USA, 2002:1-16.

- [87] Jeffrey Dean ,Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters[C]. In OSDI, 2004:137-150.
- [88] Ghemawat S, Gobioff H, Leung S T. The Google file system[C]. ACM SIGOPS operating systems review. ACM, 2003, 37(5) : 29-43.
- [89] 崔星灿,禹晓辉,刘洋,吕朝阳. 分布式流处理技术综述. 计算机研究与发展[J]. 2015, 52(2) : 318-332.

众包数据管理系统

- [90] Thomas C Redman. The impact of poor data quality on the typical enterprise[C]. ACM, 1998,41(2) :79-82.
- [91] Institute of Medicine. To err is human , building a safer health system[J]. National Academy of Sciences. 1999.
- [92] W Eckerson. Data quality and the bottom line: Achieving business success through a commitment to high quality data[R]. The Data Warehousing Institute, 2002.
- [93] <http://tech.163.com/11/0823/01/7C3S1GKG000915BF.html>.
- [94] <http://tech.qq.com/a/20110906/000171.htm>.
- [95] Jiannan Wang, Tim Kraska, Michael J. Franklin, Jianhua Feng: CrowdER: Crowdsourcing Entity Resolution[J]. PVLDB,2012,5(11) :1483-1494.
- [96] Ju Fan, Meiyu Lu, Beng Chin Ooi, Wang-Chiew Tan, Meihui Zhang: A hybrid machine- crowdsourcing system for matching web tables[J]. ICDE,2014:976-987.
- [97] Michael J Franklin, Donald Kossmann, Tim Kraska, Sukriti Ramesh, Reynold Xin. CrowdDB: answering queries with crowdsourcing[C]. SIGMOD Conference,2011 : 61-72.
- [98] Adam Marcus , Eugene Wu, David R Karger, Samuel Madden, Robert C. Miller: Demonstration of Quirk: a query processor for humanoperators[C]. SIGMOD Conference,2011 :1315-1318.
- [99] Hyunjung Park, Richard Pang, Aditya G Parameswaran, Hector Garcia- Molina, Neoklis Polyzotis, Jennifer Widom: Deco: A System for Declarative Crowdsourcing[J]. PVLDB,2012,5(12) :1990-1993.
- [100] Xuan Liu, Meiyu Lu, Beng Chin Ooi, Yanyan Shen, Sai Wu, Meihui Zhang. CDAS: A Crowdsourcing Data Analytics System[J]. PVLDB,2012,5(10) :1040-1051.
- [101] Aditya G. Parameswaran, Hector Garcia- Molina, Hyunjung Park, Neoklis Polyzotis, Aditya Ramesh, Jennifer Widom. CrowdScreen: algorithms for filtering data with humans[C]. SIGMOD,2012:361-372.
- [102] Anish Das Sarma, Aditya G. Parameswaran, Hector Garcia-Molina, Alon Y. Halevy. Crowd-powered find algorithms[J]. ICDE, 2014:964-975.
- [103] Aditya G Parameswaran, Stephen Boyd, Hector Garcia- Molina, Ashish Gupta, Neoklis Polyzotis, Jennifer Widom. Optimal Crowd- Powered Rating and Filtering Algorithms [J]. PVLDB,2014, 7 (9) : 685-696.
- [104] Adam Marcus , Eugene Wu, David R Karger, Samuel Madden, Robert C Miller. Human- powered Sorts and Joins[J]. PVLDB,2011 , 5(1) :13-24.
- [105] Hyunjung Park , Jennifer Widom. CrowdFill: collecting structured data from the crowd[C]. SIGMOD, 2014 :577-588.
- [106] Stephen Guo, Aditya G Parameswaran, Hector Garcia-Molina. So who won?: dynamic max discovery with the crowd[C]. SIGMOD, 2012 :385-396.

- [107] Adam Marcus, David R. Karger, Samuel Madden, Rob Miller, Sewoong Oh. Counting with the Crowd [J]. PVLDB, 2012, 6(2) :109-120.
- [108] <http://crowdflower.com/>.
- [109] P G Ipeirotis, F Provost, J Wang. Quality management on amazon mechanical turk [R]. In SIGKDD Workshop on Human Computation, 2010: 64-67.
- [110] Tingxin Yan, Vikas Kumar, Deepak Ganesan. CrowdSearch: exploiting crowds for accurate real-time image search on mobile phones[R]. MobiSys, 2010: 77-90.
- [111] M S Bernstein, G. Little, R C Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, K. Panovich. Soylent: a word processor with a crowd inside[C]. UIST, 2010: 313-322.
- [112] A G Parameswaran, A D Sarma, H Garcia- Molina, N Polyzotis, J Widom. Human-assisted graph search: it's okay to ask questions[J]. PVLDB, 2011, 4(5) :267-278.
- [113] 刘云浩. 群智感知计算[J]. 计算机学会通讯, 2012(10).
- [114] 丁宇, 车万翔, 刘挺, 张梅山. 基于众包的词汇联想网络的获取和分析[J]. 中文信息学报, 2013, 27(3) :100-106.
- [115] 张志强, 逢居升, 谢晓芹, 周永. 众包质量控制策略及评估算法研究[J]. 计算机学报, 2013, 36(8) : 1636-1649.
- [116] Yudian Zheng, Jiannan Wang, Guoliang Li, Reynold Cheng, Jianhua Feng. QASCA: A Quality-Aware Task Assignment System for Crowdsourcing Applications[C]. SIGMOD, 2015.
- [117] Ju Fan, Guoliang Li, Beng Chin Ooi, Kian-lee Tan, Jianhua Feng. iCrowd: An Adaptive Crowdsourcing Framework[C]. SIGMOD, 2015.
- [118] Susan B Davidson, Sanjeev Khanna, Tova Milo, Sudeepa Roy. Using the crowd for top-k and group-by queries[C]. ICDT, 2013: 225-236.
- [119] Jiannan Wang, Guoliang Li, Tim Kraska, Michael J. Franklin, Jianhua Feng. Leveraging transitive relations for crowdsourced joins[C]. SIGMOD Conference, 2013: 229-240.
- [120] Xi Chen, Paul N. Bennett, Kevyn Collins- Thompson, Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting[C]. WSDM, 2013: 193-202.
- [121] Busa-Fekete R, Szorenyi B, Cheng W, et al. Top-k selection based on adaptive sampling of noisy preferences[C]. ICML, 2013: 1094-1102.
- [122] A R Khan, H Garcia- Molina. Hybrid strategies for finding the max with the crowd[R]. Technical report, 2014.
- [123] P Ye, U EDU, D Doermann. Combining preference and absolute judgements in a crowd-sourced setting [C]. ICML'13Workshop: Machine Learning Meets Crowdsourcing, 2013.
- [124] P Venetis, H Garcia-Molina, K Huang, N Polyzotis. Maxalgorithms in crowdsourcing environments[C]. WWW, 2012: 989-998.
- [125] Gianluca Demartini, Djellel Eddine Difallah, Philippe Cudré- Mauroux. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking[C]. WWW, 2012: 469-478.
- [126] Yan Yan, Rómer Rosales, Glenn Fung, Jennifer G. Dy. Active Learning from Crowds[C]. ICML, 2011: 1161-1168.
- [127] Jianhong Feng, Guoliang Li, Henan Wang, Jianhua Feng. Incremental Quality Inference in Crowdsourcing[C]. DASFAA (2), 2014: 453-467.

- [128] Enrique Estellés Arolas, Fernando González- Ladrón- de- Guevara. Towards an integrated crowdsourcing definition [J]. Science, 2012, 38(2) : 189-200.

在线数据分析与管理系统

- [129] Sergey Melnik , Andrey Gubarev, Jing Jing Long , Geoffrey Romer , Shiva Shivakumar, Matt Tolton , Theo Vassilakis. Dremel: Interactive Analysis of Web- Scale Datasets [C]. Proceedings of the 36th International Conference on Very Large Data Bases , 2010: 330-339.
- [130] <http://drill.apache.org>.
- [131] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, Ion Stoica. BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data [C]. ACM EuroSys, 2013.
- [132] Sameer Agarwal, Aurojit Panda, Barzan Mozafari, Anand P. Iyer, Samuel Madden, Ion Stoica. Blink and It's Done: Interactive Queries on Very Large Data [J]. PVLDB, 2012, 5(12) : 1902-1905.
- [133] Online Aggregation, Joseph Hellerstein, Peter Hass, Helen Wang, SIGMOD 97 [C]. SIGMOD , 1997: 171-182.
- [134] <http://www.impala.io>.
- [135] Michael Armbrust, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K. Bradley, Xiangrui Meng, Tomer Kaftan, Michael J. Franklin, Ali Ghodsi, Matei Zaharia. Spark SQL: Relational Data Processing in Spark [C]. SIGMOD Conference, 2015: 1383-1394.
- [136] Yingjie Shi, Xiaofeng Meng, Fusheng Wang, Yantao Gan. You can stop early with COLA: online processing of aggregate queries in the cloud [C]. ACM, New York, NY, USA, 2012: 1223-1232.

其他

- [137] Shweta Dinesh Bijwe, Prof P L Ramteke. Database in Cloud Computing - Database-as-a Service (DBaaS) wth its Challenges [J]. International Journal of Computer Science and Mobile Computing, 2015 (2) : 3-4. i.
- [138] D H Manjaiah, B Santhosh, Jeevan L J Pinto. BigData: Processing of Data Intensive Applications on Cloud [J]. Computational Intelligence for Big Data Analysis, 2015 (19) : 201-217.
- [139] Dr. Muhammad Awais Shibli. Security-as-a-Service for Column Oriented NoSQL Databasesin Cloud [D]. 2015(2).
- [140] <http://aws.amazon.com/es/ec2/>.
- [141] Sivasubramanian S. Amazon dynamoDB: a seamlessly scalable non- relational database service [C]. SIGMOD , 2012: 729-730.
- [142] Chang F, Dean J, Ghemawat S, et al. Bigtable: A distributed storage system for structured data [J]. ACM Transactions on Computer Systems (TOCS), 2008, 26(2) : 4.
- [143] Krishnan S P T, Gonzalez J L U. Google Cloud SQL [M]. Apress, 2015: 159-183.
- [144] Gusev M, Ristov S. Superlinear speedup in Windows Azure cloud [C]. IEEE, 2012: 173-175.
- [145] Corbett J C, Dean J, Epstein M, et al. Spanner: Google 's globally distributed database [J]. ACM Transactions on Computer Systems (TOCS), 2013, 31(3) : 8.
- [146] 杨传辉. 淘宝 Oceanbase 云存储系统实践 [J]. 程序员, 2011(5) : 78-80.
- [147] AnHai Doan, Alon Halevy, Zachary Ives. 数据集成原理 [M]. 孟小峰, 马如霞, 马友忠, 等. 北京: 机械工业出版社, 2014.
- [148] 孟小峰, 刘伟, 姜芳芳, 等. Web 数据管理: 概念与技术 [M]. 北京: 清华大学出版社, 2014.

作者简介

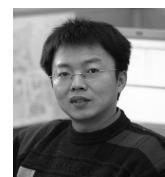
孟小峰 博士，中国人民大学信息学院教授，博士生导师。中国计算机学会会士（2013）、常务理事（2011-）、数据库专委会秘书长（1999-）、《Journal of Computer Science and Technology》、《Frontiers of Computer Science》、《软件学报》、《计算机研究与发展》等编委。近期主要研究领域为数据库系统与数据管理，包括数据融合与知识融合、面向新型硬件大数据管理、大数据隐私管理、大数据分析、以及交叉性研究如社会计算等。Email: xfmeng@ruc.edu.cn。



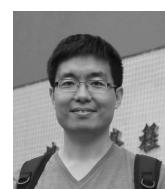
高军 博士，北京大学信息科学技术学院，教授，博士生导师，数据库专委会委员。主要研究方向：分布式数据管理、图数据管理等。Email: gaojun@pku.edu.cn。



崔斌 博士，北京大学信息学院，研究员，博士生导师，中国计算机学会杰出会员。主要研究方向：数据库、数据挖掘等。Email: bin.cui@pku.edu.cn。



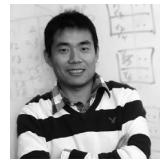
禹晓辉 博士，山东大学计算机科学与技术学院教授，博士生导师。中国计算机学会数据库专委会委员、大数据专家委委员，《Information Systems》编委。近期主要研究领域为大数据管理与分析，包括分布式流数据管理、时空数据挖掘、社交媒体数据分析等。Email: xyu@sdu.edu.cn。



李国良 博士，清华大学计算机系，副教授，中国计算机学会会员，数据库专委会委员。主要研究方向：数据库、群体计算、时空数据处理等。Email: liguoliang@tsinghua.edu.cn。



李飞飞 博士，美国犹他大学计算机系，副教授。主要研究方向：数据库系统，大数据管理理论及系统设计开发，以及云数据管理的安全性等。Email: lifeifei@cs.utah.edu。



陈刚 博士，浙江省大数据智能计算重点实验室主任，软件所副所长，教授，博士生导师，数据库专委会委员。研究方向集中于数据库、大数据分析领域，研究成果应用于大型互联网系统、国家网络安全重点工程、大型电信运营商等领域。Email: cg@zju.edu.cn。



王秋月 博士，中国人民大学信息学院讲师。主要研究方向为数据库与信息系统、信息检索、知识库、自然语言问答等。E-Mail: qiuyuew@yahoo.com。



经验软件工程的挑战和发展趋势

CCF 软件工程专委会

摘要

软件系统在当今世界已经有极其广泛的应用，大量新型的软件相关系统在不断地被开发和使用，譬如社会软件服务系统、软件生态系统等。与此同时，开放和自治的软件开发环境正在彻底变革软件的开发方法、过程。如何处理、理解和利用软件开发中产生的大量数据，使之可以促进新环境下软件质量的管理与改进，是我们面临的严峻挑战。经验软件工程正是这样一门学科，其研究内容涉及如何在软件工程的环境中获得经验数据、如何进行数据分析，从而提出假设并进行科学求证，最终指导软件过程和质量的改进。

关键词：经验软件工程，软件工程，数据分析

Abstract

Nowadays, software systems are being developed and applied in an increasingly wide scope. There are various emerging types of software systems such as Social Software Service Systems and Software Ecosystems. On the other hand, open and autonomy software environments have changed the software development methodologies and processes. How to process, understand, and leverage the data produced in the software development process, facilitate the management and improvement of software quality is quite a challenge. Empirical Software Engineering is such a discipline, which researches and provides some systematic methods to get data from the environment of software engineering, analyze the data, formulate the hypothesis and evaluate the results. Finally, apply the outcome to improve the software processes and products.

Keywords: Empirical software engineering, software engineering, data analysis

1 引言

任何学科的发展都要依赖于对这个学科所要解决的基本问题的理解。这就涉及对该学科应用领域的建模，以及对问题的解决过程和解决效果的建模。对软件工程而言，不仅要对软件产品的各种特征建模，譬如可靠性、可移植性、效率等，还要对软件项目的各种特征，譬如成本、进度等进行建模。而更重要的是，还要理解开发过程和产品特征之间的关系，例如什么样的开发过程在什么条件下可以产生什么样的产品特征。也就是说，我们不能脱离工程的要素去孤立地看待产品质量问题。

每个学科解决问题的能力都会随着领域经验的提高而提高。其基本方法是把经验封装到模型中，并基于实验、经验证据、反馈来验证和确认模型的正确性。对知识的封装

可以让我们站在更高的抽象层次上理解我们的问题空间、解决空间，并通过应用反馈和学习让我们知道哪些方法行之有效。

这是一种在很多领域都适用的方法，譬如物理、医学、制造业等。以物理为例，它瞄准的是对物理世界的理解，并将研究分为理论研究和实验研究两大阵营，物理学科的进步正是这两大阵营相互影响的结果。理论研究者建立模型来解释世界，预测那些可以度量的事件所可能导致的结果，这些模型既可以基于理论，也可以基于之前的实验数据。实验研究者则通过观察和度量，来证实或者推翻理论，或者探索新的理论。这是一个建模、实验、学习和再建模的闭环。

与上述学科一样，软件工程也需要这样高层次的方法来推进学科知识的发展。软件工程也是一门实验科学，需要建模、实验、协作来理解各种过程和技术工作的状态、制约的因素以及改进的机会。我们也必须从应用中学习并改进自身对软件工程世界的理解。

但软件工程是面向人的学科，与其他工程学科相比，人的智力活动在软件开发中起着决定性作用。人作为软件开发者和工程实践者，通过与软件工程技术进行交互创造出软件系统。因此，在软件工程的研究中，人的因素和影响往往扮演着更重要的角色。没有任何两个软件是相同的。过程是变量，目标是变量，环境也是变量。我们需要建立、分析和评价软件过程、产品以及相关环境的各种有用模型，以及模型之间的交互作用。经验软件工程（Empirical Software Engineering）就是这样一种方法，通过实验手段仔细分析、验证假设以及适用的环境和条件，从而使得在小范围内揭示和确认的方法可以在更大、更复杂的环境中得到应用。

在经验软件工程研究中，研究者和实践者的共生关系相对于其他学科更为明显。研究者的任务是理解开发过程和产品的自然属性以及它们之间的关系；实践者的任务是利用这些知识建立改进系统。研究者需要的实验室中通常有实践者建立的软件系统，而实践者则需要研究者提供的模型去建立更好的系统。

经验软件工程的起源可以追溯到 20 世纪 70 年代 Weinberg 和 Ticky 的工作。他们提出并鼓励采用科学的方法来细致地观察和评鉴软件开发方法。最早的实验出现在 20 世纪 80 年代早期，Victor Basili 在 Maryland 大学和 NASA 的软件工程实验室的团队，使用实验的方法来检查软件工程技术的适用性。

进入 21 世纪后，随着互联网技术的发展，我们可以看到软件的规模、复杂度在戏剧化地攀升，系统中的系统（System- of- Systems）、社会软件服务系统（Social- Software-Service-System）、软件生态环境（Software- Ecosystem）等新的软件形态和社区正在形成。2013 年 9 月，Gartner 在对 2014 年具有战略意义的十大技术与趋势做出总结的时候，首次引入了“软件定义一切（Software Defined Anything）”的概念。事实上，软件定义数据中心、软件定义网络、软件定义系统等正在成为业界共识，软件的地位变得越来越突出。这种发展也意味着承载其服务的软件的需求爆炸性增长，传统以公司为主体的软件开发模式已经无法满足超量的软件需求，市场呼唤一种快速、低廉并且高质量的软件开发方法——开源/开放的软件开发环境。开放环境的协同模式改变和颠覆了软件业的工作方式，真正释放了软件开发人员的创造力与生产力。从业者通过自愿、兴趣和协作的方式，

依托互联网的交互平台形成软件的开发和贡献群体。自由、开放地交流沟通、共享经验、参与协作，给软件的开发活动融入了更多的社会学元素，也带来了更多的不确定因素，使得软件工程这个社会-技术系统变得更加复杂。其突出表现在以下几个方面。

1) **复杂社会-技术系统中的大数据**。传统软件开发环境中已然存在大量的数据，譬如各类工程文档和过程管理数据，一直存在的问题都是数据多、质量低，缺少使用的方法。因此很多的数据都是束之高阁，无人也无法使用。在开放环境下，预定义、有结构的开发方式被逐渐摒弃，文档少了、评论多了，数据的噪声也更大了。人们在畅快淋漓地写代码的同时，发现软件的维护和演化变得如此困难，到底该做什么？怎样才能给我们带来最大的价值？如何评估风险？究竟该推荐什么人才能高质量地完成任务并保持社区的繁荣？没有确定的过程可以遵循，没有明确的资源可以使用，没有合同约束你该做什么，这就是开放环境带给我们的挑战。经验软件工程的核心思想就是在观察和推理的基础上探索理论。开放环境给软件工程带来了挑战，同时也带来了机遇。观察开发过程中的数据、分析数据间的关系，进而推理出合适的理论，这一过程有助于我们在软件开发活动中开展评估、选择和决策工作。

2) **软件开发的效率和质量**。开放环境中的软件涉众从封闭、有限的群体转变成开放、无限的群体，社区内的成员可以自由、开放地交流沟通、共享经验、参与协作，软件的开发活动融入了更多的社会学元素。虽然利用大众的群体智慧为快速、低廉地开发高质量软件提供了可能，但生产方式的变革要求我们必须重视社会学形态对软件开发的方式、过程、生产率和质量的影响。譬如开源软件由于其松散的组织方式、自由的开发贡献和无承诺的交付方式，使得大量的开源软件存有各种缺陷和安全隐患。而这些问题又由于生产方式的变革，使得传统的方法难以发挥作用。因此，封闭环境中的效率和质量的提高依赖于过程和方法的改进，而在开放环境下，由于需求、资源和边界的不确定，这要求我们不得不考虑更加柔性、多变的改进方法。譬如准确的资源推荐而不是分配，基于大众使用评论的质量评价和改进等，这也给经验软件工程带来了更多的应用和发展空间。

3) **开放社区中的群体智慧**。开放社区建立了一种自治合作的软件生态环境，不仅是软件的开发者，软件的用户也以各种方式参与其中，形成了一种多层次、多元化的软件涉众群体。在这种自治的软件生态环境中，社会学的属性会映射到个人和群体的行为中，传统的软件过程方法已经不适应这种社会自治的合作方式，大众在竞争与协作中形成的各种知识必然会以某种方式促生新的群体智慧。如何理解开放环境下群体智慧的复杂性，如何建立支持开放、自治的社区合作软件开发模式，如何适应群体合作、演化和发展的软件过程方法，以及基于群体智慧的需求分析、代码理解、缺陷预测/定位/修复等如何支持软件质优价廉的开发，这些用经验软件工程的方法都可以找到合适的解决方案。

因此，我们看到，经验软件工程从最初的软件工厂开始发展，今天在软件生产方式上有重大变革的时机，其基于观察、实验和理论推理的方法对不确定的软件开发环境中新问题的分析、理解和改进提供了有效的解决途径，特别是针对软件工程中的大数据与基于大数据的知识获取、分享、运用，已经有了非常重要和前瞻性的研究，并且将推动软件工程领域的技术发展。

2 经验软件工程的基本研究方法

Empirical 在英文中是“经验的”或“实证的”意思。因此，Empirical Software Engineering 被译为“经验软件工程”或“实证软件工程”。但在英文中 Empirical 区别于 Experiential，这里的“经验”不单单是人在实践中的主观体验和认知，更强调从实践中获得的客观证据。

经验软件工程的研究内容涉及如何获得经验数据、如何进行数据分析，从而提出假设并进行科学求证，最终指导实践活动。根据数据形式的不同，经验研究又可分为定性研究（Qualitative Research）和定量研究（Quantitative Research）。定性研究中处理的数据主要为文字、图像等难以量化的信息，其研究方法主要来自社会科学领域，如编码（Coding）、主题归纳（Thematic Synthesis）、持续比较（Constant Comparison）、扎根理论（Grounded Theory）等。定量研究中处理的是能够被量化的数据，会用到概率统计（如假设检验、回归建模）、仿真试验、机器学习等方法。通常根据数据获取的方式、研究对象、分析方法、证据强度等的不同将这些研究方法分为受控实验（Controlled Experiment）、调研（Survey Research）、案例研究（Case Study）和行动研究（Action Research）等，这些方法虽然各有特点，但并不排斥，也常常混合使用。比如通过案例研究发现可能的因果关系，再通过更广泛的调研或设计受控实验进一步求证。在关于相似问题的大量第一手经验研究的基础上，还可以使用系统评价（Systematic Review）开展二级经验研究。

经验研究是以改进软件工程实践为目的，将理论上的概念运用于软件工程实践中，还将应用经验形成知识加入软件工程系统中，以改善过程、方法和工具，验证理论与模型。在软件工程中，我们需要对理论和“一般常识”进行经验验证，判断是否存在某种关系（如可维护性及其度量因子之间的关联）需要选择模型、技术或工具，并判断它们在实践中的效果。

经验软件工程的研究目标是实施更好的经验研究、注重因果机制并生成理论，实现迭代和改进。用良好的经验能够更迅速地构建软件工程知识体系，快速删除低回报的想法，识别和评估高回报的想法，探究重要的实际想法。经验软件工程将受控实验、调研、案例研究等研究方法应用到软件工程研究中，积累软件工程领域的经验知识。

下面我们就软件工程中的主要经验研究方法和需要考虑的问题做一些简介。

2.1 受控实验

受控实验是一种实验性研究方法，它针对一个或多个可验证的假设，并对其中相互独立的自变量（Independent Variable）加以控制，系统地操控实验条件，尽量消除非研究自变量的影响，并观察和验证被研究的自变量是如何影响某些可度量的结果的，即因变

量 (Dependent Variable)。

大部分受控实验的主要目的是验证某个假设或者关系。假设检验一般是针对前者，后者主要是基于所收集的数据建立一个关系模型（如预测模型），这个模型通常可以用多元统计方法得到，例如回归技术，然后用实验进行评价。

实施一个实验涉及几个不同的步骤。这些步骤是：确定范围、计划（包括实验人员、实验环境、实验假设的阐述，确定实验变量，实验方法和过程的设计等）、操作（包括实验准备、实验执行和数据验证等）、分析和解释、展示和打包（如论文发表等）。

2.2 调研

调研的目的是为了给出与特定研究问题相关的观点的全面总结，一般通过收集来自于人或者与人有关的信息，来描述、比较或者解释人们的知识、态度和行为。在实际应用中，基于调研的方法有3种形式，即访谈、问卷、文献。这几种方法并不是互斥的，很多研究人员将几种方法混合使用。通过数据三角检验法往往能设计出更合理的调查，甚至得出更有说服力的研究结果。

Kitchenham 等在 2002 年发表了一系列文章[219] ~ [221]，系统介绍了调研方法的原则，这些原则也成为软件工程领域进行调研研究的执行指南。此后，调研方法在经验软件工程领域越来越多地被采用，其优点是能够直接、全面地获知受访对象的观点、经验等第一手数据。然而由于这一方法受人为因素影响较大，往往难以有效地避免访问者与被访者的个人偏见，同时由于采样方法多基于便利条件，很难保证调查结果的有效性。这些都是软件工程研究人员使用调研方法时需要考虑和关注的问题。

基于文献的调研在最近 10 年得到了快速的发展。2004 年，Kitchenham^[219] 提议软件工程研究者应当开展基于证据的软件工程 (Evidence-based Software Engineering, EBSE)，并为此正式在医学领域引入了系统文献综述方法。这是一种是针对某个特定的研究问题、主题领域或现象，识别、评估及解释与之相关的所有可用研究的手段。相对于原始研究 (Primary study)，系统文献综述是一种二次研究 (Secondary Study)。

传统综述更多的是基于专家的经验和判断，强调权威性。系统文献综述则更强调客观性，即强调分析和结论的客观性，因此要求对调研过程要有详细的描述。系统文献综述有以下特点：

- 系统文献综述要求在开始时定义评价协议，在协议中声明文章的研究问题以及分析将使用的方法；
- 系统文献综述是基于一个清晰定义的目的而尽是完整地获取相关研究的搜索策略；
- 系统文献综述会记录所使用的搜索策略，以便于读者能够认知它的准确程度和完整程度；
- 系统文献综述需要明确的收录和排除标准来评估每一篇可能的原始研究；

- 系统文献综述会列出即将从每篇原始研究中获取的信息（包括质量标准），以便于评估每篇原始研究。

2.3 案例研究

案例研究是“以典型案例为素材，并通过具体分析、解剖，对当前某一现实环境中的现象进行考察的经验性研究方法”。案例研究方法非常适用于各种软件工程研究，因为软件开发过程是一个极其复杂的过程，往往很难孤立地研究各个因素；同时基于受控实验的成本等因素，它往往难以模拟大型复杂系统，也难以考虑政策等外界因素，因而案例研究室是对受控实验的一个很好补充，它包含了很多实验无法体现的特质，比如规模、复杂度、不可预测性以及动态性。

对于多种类型的软件工程研究，只要被研究对象是当前存在的且难以单独研究的，都可以采用案例研究的方法。案例研究不会产生和受控实验一样的结果，如因果关系等，但它可以让我们对真实环境中的现象有更深入的理解。正因为与分析型、受控的经验研究不同，案例研究总被批评为价值小、无法泛化、带有研究者偏颇性等。但如果在实践中能采用合适的研究方法，并接受“知识不仅是统计学意义的”观点，则上述批判就不是问题了。

案例研究也非常适合对软件工程方法和工具在工业界的使用情况进行评估，因为它可以避免按比例放大问题。但在进行案例研究时，必须设法最小化混杂因子所带来的影响。

在软件工程中，经验研究以及它们的影响在不断增大。然而经验研究方法论大部分集中在实验研究中，因此以分析为主的研究范式不足以反映复杂的现实生活问题，尤其是涉及人类以及他们的互动技术方面的问题，为应对复杂问题，通常采用案例研究。Runeson 和 Host 给出了软件工程学科的案例研究的框架，这篇论文具有里程碑意义，之后的大部分相关论文都是根据这个框架做案例研究的。

案例研究的步骤包括：①案例研究设计（定义目标和制定计划）；②准备数据采集（定义数据采集的程序和协议）；③收集证据（介绍数据采集的过程，并收集数据）；④对数据进行分析处理并讨论；⑤撰写案例研究报告。这个过程和其他类型的经验研究几乎是相同的，然而，案例研究的方法具有灵活的设计策略，那么它在步骤上会产生大量的迭代。

案例研究是对现实世界问题进行的研究，其目的是在控制水平和现实程度之间寻找一种平衡。现实的情况往往是复杂和不确定的，因而案例研究具有高度的现实主义。案例研究的优势在于它更容易设计并且更加真实，劣势则体现在它的结果难以归纳，也更难解释。

2.4 经验数据的处理和分析方法

在经验研究中，需要通过对调研、案例及实验中的数据做出分析，从而得到一些结

论。经验研究中常见的对数据处理的方法包含：数据预处理、数据统计分析、假设检验及数据分类和预测。

数据预处理是为了提高数据处理的质量，预先对数据中的虚假数据点等进行过滤处理，以提高后续分析的质量和效果。数据预处理的方法有多种：数据清理、数据集成、数据变换、数据规约等。数据清理通过填写缺少的值、光滑噪声数据、识别或删除离群点并解决不一致性来“清理”数据，从而达到数据格式标准化、异常数据清除、错误纠正及重复数据清除的目的。数据集成则是将多个数据源中的数据结合起来并统一存储。数据变换则通过平滑聚集、数据概括、规范化等方式将数据进行转化。数据规约可以通过规约表将数据的量减少但仍能基本保证数据的完整性，其结果与规约前的结果相同或几乎相同。

数据统计分析是指运用统计方法及分析有关的知识，结合定量与定性，进行研究活动。数据统计分析需要描述所分析的数据性质、研究群体的数据关系，创建模型，总结数据间的关系。其目的是分析出差异、趋势，查找问题。常用统计分析包含对数据集中趋势、离散程度、数据间依赖关系的度量，以及对数据的图形可视化。

假设检验是数理统计学中根据一定假设条件由样本推断总体的一种方法。假设检验的目的是确定是否能够基于某统计分布的一个样本拒绝某个空假设 H_0 。检验可以分为参数检验和非参数检验。常用的假设检验方法有： t -检验 (t -test)、曼-惠特尼检验 (Mann-Whitney)、 F -检验 (F -test)、配对 t -检验 (Paired t -test)、威尔科克森检验 (Wilcoxon)、符号检验 (Sign test)、方差分析检验 (ANOVA)、克鲁斯卡尔-沃利斯检验 (Kruskal-Wallis)、卡方检验 (Chi-square) 等。

数据分类和预测则是从以往的数据中获取数据分类，从而预测、应对未知的东西，主要是使用机器学习来实现分类和预测。分类预测的算法根据数据功能和形式的类似性开展工作，常用的算法有：决策树学习法、贝叶斯方法、基于核的算法、人工神经网络、集成算法等。

2.5 经验分析的质量与实验有效性验证

除了上面介绍的经验研究的主要方法和技术，在软件工程中开展经验研究还涉及一些相关的重要问题，这在几乎所有的经验研究中都要考虑到。

2.5.1 数据质量分析

软件开发数据的快速积累为软件工程研究者提供了巨大机遇，使软件工程研究者得以基于大量数据对研究问题进行定量研究。与此同时，越来越多的机器学习、数据挖掘算法被引入软件工程研究中，以帮助软件工程研究者更好地理解、利用软件开发数据。然而，软件开发数据自身的质量却没有被很好理解。这些数据可能由于项目的实践不同、工具使用者的使用方式不同，并不能忠实地反映软件开发的实际过程或并不符合软件工

程研究者对其所做的假设。例如一个项目在其演变过程中会更换版本控制系统，造成数据不完整等问题。若软件开发数据的质量存在问题，由这些数据所得到的结论的可信度将会受到影响。

目前，软件开发数据中的质量问题受到了越来越多研究者的关注。首先，对数据分析方法的研究是最近的一个趋势，其次，探讨数据本身存在的问题及其对研究和实践的影响也正在逐步成为一个重要方向。

2.5.2 经验证据的强度

经验软件工程中的一个重要问题是：在软件工程领域的原始研究的质量通常不高而使用的研究方法各异的情况下，我们该如何对已有的经验研究进行恰当的评价和比较。为此，证据的强度便成了一个很重要的属性。

目前存在着一些能对证据的强度做出判断的系统，其中大部分都认为随机性的实验属于高等级，而观察性的研究（如案例研究）和研究者的个人观点（如调研）则属于相对低等级。不过这也存在着一个固有的缺陷：随机性的实验并不是在所有的研究中都可行，而在某些情况下，观察型的研究反而能够提供更好的证据。

通常来说，以下因素会降低证据的强度：研究质量的限制、结果的不一致、相关程度的不可靠性、不严密或者数据稀少、报告中的偏差。而可以增加证据强度的因素有：联系的强或者非常强的证据、应梯度（Dose- Response Gradient）的出现、对影响的低估等。

2.5.3 经验研究的有效性问题

在经验软件工程中，研究者首先要了解的问题就是经验研究结果的有效性（Validity），以及影响有效性的不利因素，即有效性威胁有哪些。有效性具有推断的性质，每个经验研究结果的有效性都面临着一定范围的威胁。如同风险管理，研究者需要知道威胁有效性的相关因素有哪些，以便适当地控制或者接受这些威胁。为了帮助识别和处理这些威胁，Wohlin 等人把经验软件工程中的有效性威胁主要分为 4 种类型：结论有效性、内部有效性、构造有效性和外部有效性。

1) 结论有效性 (Conclusion Validity)：如何确保我们在一个经验研究中使用的方法、策略确实与我们观察到的实际结果有关。也就是说，这种结果是否具有统计意义。

2) 内部有效性 (Internal Validity)：我们如何应用这种方法得到需要的结果，当然，也可能有一些不受控制的因素会导致实际的结果。

3) 结构有效性 (Construct Validity)：经验研究背后的理论与观察结果之间的关系。即使我们已经建立了实验方法和观察结果之间的因果关系，其方法也有可能不受我们控制，以至于结果偏差很大。

4) 外部有效性 (External Validity)：我们是否可以将结果的概括总结应用到其他的研究领域。因为即使已经确定我们的研究结果具有统计意义，也不能轻易下结论说这种

方法具有普适性。

2.5.4 经验研究的复制

复制（Replication）是一种经验软件工程研究中非常常见的研究形式。复制研究旨在帮助研究者建立、完善某个特定研究主题中结果和相关上下文的系列知识 [Basili99]。在经验软件工程的研究（特别是实验）当中，实验的上下文环境（Context）往往会对实验的结果产生很大的影响，因而，考察某个软件工程实践、方法以及工具在不同的上下文中是否仍然有效就成为经验软件工程研究的主要目标之一。复制是实现该目标的有效手段。通过复制研究，不管发现的结果与原始研究是否一致，复制研究都被认为是对原始研究结果的有效补充。Shull 等人将复制研究分成两种，一种是严格复制（Exact Replication），另外一种是概念复制（Conceptual Replication）[Shull08]。前者要求尽量保持复制研究中的上下文与原始研究中的上下文一致或者接近，后者则仅仅只是针对相同的研究问题，而上下文环境，甚至实验过程都可以不同。

3 软件工程各领域的主要经验研究成果

我们在软件工程学科当中选择了经验研究较多的几个领域，对经验研究在这些领域中的主要有代表性的成果进行总结归纳。

3.1 估算与预测

根据文献 [101] 的总结，软件成本估算相关的研究最早可以追溯到 20 世纪 60 年代。经过近 50 年的发展，软件成本估算技术取得了长足的发展，但成本估算仍然是目前软件项目管理中的主要难题之一。在软件工程实践中，“软件成本”是一个相对模糊且宽泛的概念。根据应用场景的不同，软件成本估算的目标可能涵盖货币成本、工作量、资源、进度等多个方面。考虑到软件开发的高智力特性，人员成本一般占整个软件项目成本的绝大多数，因此在很多情况下可以用“工作量”（Effort）来代替“成本”的概念。在软件成本估算的相关研究中，大多也将“工作量估算”视为成本估算的目标。

预测是经验软件工程中另一个重要的研究领域，预测主要是研究未来的一种倾向或者趋势，譬如某些模块是否更容易出现缺陷，某些人修复缺陷的质量高等等。

在经验软件工程中，估算和预测研究的主要方面有以下几类。

3.1.1 成本估算

软件成本估算研究发展至今，已有大量估算方法、模型被提出和使用。根据使用的估算方法不同，软件成本估算主要可分为专家估算、基于类比的估算、基于参数化模型的估算、基于机器学习技术的估算和组合估算方法等^[87,96,126,23,26,190,193,82-82]。

近年来，针对软件成本估算领域，很多研究人员针对软件成本估算的不同方面给出了系统评价。

2005 年，为了确定在软件工程成本估算中经验研究结果的一致性，文献 [114] 利用预先建立的搜索准则，构建了一个详尽的文献检索，研究结果表明 25% 的研究结论是不确定的。相关研究者建议在以后的研究中需要关注更细节的问题。

2007 年，文献 [101] 中全面回顾、分析了软件成本估算的各种代表性方法，也归纳、讨论了与成本估算强相关的软件规模度量问题，进一步研究了软件成本估算方法的评价标准，并给出了一个应用实例及其分析，最后从估算模型、估算演进、估算应用、估算内容、工具支持和人力因素 6 个方面，指出了软件成本估算方法在下一步的主要发展趋势。

在使用 COCOMO II 等参数化模型时，常需要使用本地数据对模型参数进行校准。局部校准被认为是使用参数化模型的最佳实践之一^[132]。然而，由于本地数据的积累要耗费很长的时间，消耗大量的数据收集和维护成本，所以在很多情况下需要借助其他组织的项目数据进行成本估算，即进行跨组织估算。目前在跨组织估算和同组织估算的性能比较方面存在争议，尚无统一定论。Kitchenham 等人的对比研究表明，同组织估算的性能显著高于跨组织估算^[89-90]。Jeffery 等在 ISBSG 数据集上进行的实验结果也支持这一结论^[79]。类似的研究还有 Mendes 等使用 web 开发项目的数据进行的对比实验^[121]。与之相反，Wieczorek 和 Ruhe 等发现跨组织估算的性能与同组织估算的性能无显著差异^[195]。Mendes 等人的另一项实验也表明在部分数据集上使用多组织数据和同组织数据建立的成本估算模型的性能差异不明显^[122]。

另有部分研究关注如何提高成本估算数据的可用性。如 Kitchenham 通过分析数据集的不均衡性，帮助解释 COCOMO II 模型在最初校准时的困难^[88]。Liu 等提出了一个针对软件工作量估算数据的预处理框架，用于识别和剔除异常数据点^[106]。

3.1.2 工作量估算

工作量估算也是软件项目开发的重要环节，也是制定项目开发计划的依据。现在很多经验研究是针对工作量估算的。

2005 年，文献 [167] 关注于项目早期的特征子集选取和工作量估算。相关研究者提出了一个使用灰色系统理论中的灰色关联分析的新颖方法，以解决小样本不确定性。他们采用灰色关联分析进行特征子集的选取和工作量的估算。实验结果显示该方法优于其他机器学习方法，有很大的潜力。

2006 年，文献 [168] 使用基于关联规则挖掘的方法预测缺陷的关联性和缺陷修正工作量。他们把该方法应用在超过 15 年的 200 个项目中，结果表明对于缺陷关联性预测，它的精确度很高，假阴性比率很低；对于缺陷修正工作量，缺陷隔离工作量预测和缺陷修正工作量预测的准确率都很高，并且和别的方法相比，准确度提高了至少 23%。

2007 年，文献 [113] 调研了 1995 ~ 2005 年间工作量相关的研究，从中抽取了 10 篇相关的研究。分析结果表明由于不同研究使用了不同的数据集和不同的实验设计，导致

了结果的异质性。

2009 年, 文献 [191] 提出了一种基于灰色模型 GM (1, 1) 和 Verhulst 的方法去预测软件物理时间的阶段工作量。在大型的现实世界的软件工程数据集上验证该方法, 并且将其与线性回归方法和卡尔曼滤波方法相比, 结果显示它的精确度至少分别提高了 28% 和 50%, 从而表明了该方法具有一定的潜力。

2010 年, 文献 [67] 调研了中国工业界中影响开发工作量的因素。他们在 140 个软件组织的 999 个项目中验证了 6 个因素 (项目规模、团队规模、工期、开发类型、业务范围、编程语言) 对开发工作量的影响。结果表明这 6 个因素对开发工作都有影响。结合这些影响可以更好地对项目进行管理。

3.1.3 规模估算

目前主要的软件规模度量方式有代码行、功能点、对象点和用例点 4 种^[101]。以下分别对基于这 4 种度量的软件规模估算的经验研究进行介绍。

1) 代码行 (Lines of Code, LOC) 是一种对软件规模最简单最直接的一种度量, 在软件成本估算中被大量使用^[4, 138, 143]。根据度量对象的不同, 代码行度量可分为逻辑代码行度量和物理代码行度量两种^[95]。目前在软件成本估算中使用较多的是逻辑代码行。概念数据模型在信息系统的需求分析中被广泛应用, 信息系统中的很多特征都可以在概念数据模型中体现, 文献 [178, 179] 使用概念数据模型来估算代码行, 并且在开源系统和工业界验证了该方法的有效性。

2) 不同于代码行度量, 软件功能点 (Function Point) 度量方法主要基于用户需求对软件规模进行估算, 估算结果独立于编程语言类型^[4]。目前应用较为广泛的一种功能点计算方法是国际功能点用户协会提出的 IFPUG 方法^[73]。其他的功能点度量方法还包括 Mk II FPA 方法^[187]、NESMA 方法^[133]和 COSMIC-FFP 方法^[188]等。功能点主要基于软件的功能需求来估算产品的规模, 缺乏对软件系统复杂度的考虑。此外, 功能点方法的适用范围有限, 一般认为其比较适合管理信息系统, 而对于逻辑复杂的科学计算软件等不太适用。

3) 对象点 (Object Point) 方法与功能点方法的原理相似, 通过统计需要耗费大量工作量的制品来估算软件规模, 如服务器数据表的数目、客户数据表的数目、报表和屏幕视图中的可重用比例等^[13]。

4) 用例点 (Use Case Point) 方法是随着统一建模语言 (Unified Modeling Language, UML) 在软件开发中逐渐普及而兴起的另一种软件规模度量方法, 目前在成本估算中也得到了广泛的应用。用例点方法的典型代表是 Karner 于 1993 年提出的 UCP 方法^[85]。

3.1.4 软件维护预测

为了增加新功能、修复缺陷或者适应新环境, 软件需要不断地更改, 从而会变得越来越复杂, 维护成本越来越高。因此, 对软件维护有效方法的研究引起了越来越多研究人员的关注, 如预测维护成本^[100, 211], 评估可更改性^[32], 预测软件中可能发生更改的部分^[8]等。

更改预测可以使软件开发者和维护者更有效地利用诸如同行评审、测试和检验等资源预测面向对象软件系统中哪些模块将可能发生变更，这在软件工程领域具有重要意义，它们可以作为工作量调度依据，有助于提高软件开发和维护工作的效率。到目前为止，研究人员提出了许多通过软件历史数据信息和当前源代码信息来预测模块变更趋势的方法。在预测更改的可能性的多种方法中，使用软件度量构造更改可能性预测模型是一种广泛使用的方法^[93,65]。静态软件度量不需要运行软件就可以直接得到，因此引起了很多研究人员的关注^[100,211]。Chaumun 等^[32]提出一种变更影响模型分析在软件系统各个模块中的传播以及相互影响；Zimmermann 等^[216]提出利用版本控制系统中的数据来预测软件系统在未来的版本变更趋势；文献 [185] 提出了用一个概率方法去估算当一个功能增加或修改时更易改变的类。Sharafat 等^[160]提出一种基于模块依赖和软件历史数据信息的模块变更概率预测方法；薛朝栋等对 Sharafat 等^[160]的变更预测方法进行简化，在软件依赖关系的基础上提出了一种轻量级的模块变更概率计算方法。文献 [109] 采用统计元分析方法在 102 个 Java 系统中去探究 62 个面向对象度量元预测易出缺陷类的能力。文献 [214] 搜集了类级别上的静态代码度量元和变化数据，分析发现 80% 的代码行变化存在于 20% 的类中，然后他们使用分类技术来预测易改变的类，结果显示他们提出的分类方法对于易改变类的预测是有用的。

3.1.5 缺陷预测

作为对软件质量保障具有重要实际意义的一项活动，软件缺陷预测从 20 世纪 70 年代即开始受到关注，至今仍然是软件工程领域的一个热点研究问题^[1,199,80]。目前，学术界普遍存在的一种公认的观点，即由于软件项目的复杂性，在不能确保软件项目不产生缺陷的前提下，将关注点转向在软件发布早期对软件缺陷进行合理的预测，尽量减少经发布的软件中可能存在的缺陷，并且，在软件发布之后对用户报告的缺陷及时有效地解决，以减小由于软件缺陷给用户带来的负面体验。基于此种观点，研究人员提出了一系列方法以支持软件发布之前的缺陷早期发现和软件发布之后的有效解决，主要成果体现为以下 4 个方面：

1) 缺陷预测。这方面的研究将数据挖掘、机器学习等方法与软件产品、人员和过程度量方法相结合，利用大规模的历史软件开发和缺陷数据，通过学习已知模块或组件的度量特征与缺陷之间的潜在模式，预测新增模块或组件中可能存在的缺陷。例如，文献 [210] 利用面向对象设计度量元预测高、低严重程度的缺陷。Zimmermann 和 Nagappan^[217]利用代码片段之间的依赖关系构建社会网络并利用社会网络度量元进行代码缺陷预测。Kim 等^[91]研究了缺陷数据中的误分类数据对缺陷预测模型的性能影响。Li 和 Zhang 等^[102]将主动学习、组合学习和半监督学习的思想引入软件缺陷预测中，针对可利用的软件缺陷历史数据稀少珍贵的问题，提出了基于抽样的主动式半监督缺陷预测方法。2014 年 Yuan 等^[205]人提出了面向细粒度源代码变更的缺陷预测方法，他们采用特征熵差值矩阵分析了软件演化过程中概念漂移问题的特点，并提出一种伴随概念回顾的动态窗口学习机制来实现长时间的稳定预测。

经过几十年的研究，软件缺陷预测问题取得了长足的进展，有大量的预测方法被提出和验证。然而，有研究表明，已有的缺陷预测方法大多仅在项目内适用，即需要采用同版本或同项目历史版本的缺陷数据来进行预测模型的训练^[218]。产生这一限制的原因是传统的机器学习技术是基于数据同分布假设的，这要求预测模型的训练数据与测试数据（或待预测数据）分布一致。在某些实际的软件缺陷预测应用环境中，这种同分布的本地历史数据不可获取。为解决本地历史数据缺乏的问题，一个可能的途径是进行跨项目缺陷预测（Cross-project defect prediction），即参考外部项目的历史数据（跨项目历史数据）中的缺陷预测模型，然后将其应用到当前待预测项目上。跨项目缺陷预测问题近年来得到了大量的关注，有部分研究对跨项目缺陷预测的可行性、预测算法、数据预处理方法等进行了初步探索，并取得了不错的成绩^[111, 141, 227]。

近年来很多研究者对缺陷预测研究做了系统评价，比如2013年，文献[147]的相关作者调研了1991~2011年间106篇与缺陷预测度量元相关的文章。结果表明面向对象的度量元的使用几乎是代码度量元的两倍，其中CK度量元是最常用的；面向对象和过程度量元比规模和复杂性度量元的缺陷预测能力更强，其中过程度量元比任何静态代码度量元更能较好地预测发布后版本的缺陷。同年，文献[74]分析了哪些面向对象度量元对预测易出缺陷模块是有用的，相关研究者调研了已有的基于CK+SLOC度量元的经验研究，并从中提取出29篇相关文章。研究结果表明，耦合性、复杂性和规模度量元对易出缺陷模块的预测有较强的影响。2014年，文献[76]分析了面向对象度量元与外部质量属性（可靠性、可维护性、有效性和功能性）之间的关联性，相关研究者调研了99篇相关文章，结果表明与继承度量元相比，复杂性、内聚性、规模和耦合性度量元与可靠性、可维护性有较强的关联性。2015年，文献[115]等相关作者调研了1991~2013年间使用机器学习方法的缺陷预测论文，并从中筛选了64篇文献，同时把机器学习方法分为7类。通过分析，总结出易出缺陷模块预测中机器学习技术的使用情况，评价了其预测准确度和能力，并与统计学方法进行了对比，最后总结出机器学习方法的强项和弱项。

2) 缺陷诊断。方面的研究主要关注于在软件缺陷发生之后，如何根据提交的缺陷报告将软件缺陷及时有效地分配给相关开发人员，以进行缺陷修复。因为开源软件项目缺乏统一的开发任务管理，并且由于个人开发人员对开源项目的开发兴趣具有不确定性^[207]，致使个人开发人员不可能了解所有软件组件的依赖关系和各组件的具体代码实现。在该情况下，如何在有限信息和不确定条件下进行软件缺陷诊断已经成为一个越来越突出的问题。例如，Anvik等^[9]利用文本分类方法进行缺陷修复人推荐。Xie等^[200]等利用LDA主题模型针对不同开发人员对于修复缺陷的不同程度兴趣和专长进行建模，并在此基础上提出DRETOM方法以进行缺陷修复人推荐。文献[39]提出一种基于调用堆栈重复缺陷报告聚类的方法。文献[65]分析了一个缺陷报告是真正的缺陷还是一个过时的测试。文献[209]预测缺陷可能在多长时间内修复。

3) 缺陷定位。方面的研究主要关注于在开发人员接受缺陷修复任务之后，如何依据缺陷报告描述内容、代码变更历史、邮件交流、调试运行信息、软件功能演化、运行

日志记录和个人开发经验等，理解缺陷发生的原因，定位可能导致缺陷发生的程序代码片段。尤其是在开源软件开发由开发人员意愿驱动，缺乏明确且严格描述的需求文档和测试用例的条件下，使得一些自动化的缺陷定位和修复方法不具可行性。因此，开源软件在缺陷发生之后，很大程度上要依赖于开发人员的个人能力来理解、调试和变更代码以进行缺陷修复。于是，如何提供有效手段来辅助开发人员快速定位导致缺陷的代码、理解缺陷发生的原因并提供可能的缺陷解决方案备受是软件工程研究人员关注的另一重要问题。例如，Stacy 等^[169] 基于 LDA (Latent Dirichlet Allocation) 方法建立源代码主题模型，将源代码文件用潜在语义主题进行表示，并从缺陷报告中抽出用户查询以定位修复缺陷时所需修改的源代码文件。Moin 和 Khansari^[127] 提出了一种基于代码修改日志和缺陷报告文本分类的方法，以定位程序源代码目录级别的缺陷。Zhou 等^[213] 参考缺陷报告的文本长度和缺陷报告之间的文本相似性提出了一种基于信息检索的缺陷定位方法 BugLocator。文献 [186] 利用方法调用异常来提高基于频谱的缺陷定位技术在面向对象程序中的有效性。文献 [213] 利用信息挖掘技术预测缺陷的修复位置。文献 [58] 提出一种利用简单的用户反馈的交互式缺陷定位方法。文献 [124] 提出了一种基于聚类的策略来确定故障定位的准确性。文献 [197] 利用分段和堆栈跟踪分析来提高面向缺陷报告的缺陷定位的准确度。

4) 评审后缺陷预测。评审后缺陷预测是指，采用一定的技术和方法预测未被评审发现的缺陷。该课题的相关经验软件工程研究可以大致分为两个方向：①利用历史数据建模进行预测，其中又可以细分为跨项目 (Cross-project) 模型和项目内 (Within-project) 模型；②不依赖历史数据进行预测，目前主要采用的是在生物统计学中已经非常成熟的捕获再捕获 (Capture-Recapture, CRC) 模型。目前，该领域相关研究中绝大部分都是基于历史数据的缺陷预测技术。

基于历史数据的预测模型往往会受限于历史数据的质量和数量，因而，建立起来的预测模型有一定的局限性。例如，由于大多数的新项目并没有历史数据，所以无法使用项目内模型；而 Zimmermann 等人的研究表明，即使是同一个领域的项目，采用跨项目模型来预测缺陷也会得到糟糕的结果^[218]。

如果不依赖历史数据，无疑可以大大增加预测技术在实际软件项目评审中的应用可能。其中，捕获再捕获模型是一个值得关注的方向。然而，Liu 等人采用系统评价的研究方法综合分析了目前 CRC 模型的相关研究。其结果表明，在基于 4 个基本 CRC 模型的 18 个具体的计算算子中，并没有一个相对确定的最好算子，并且，缺乏工业领域的应用也是 CRC 模型相关经验研究当前状态存在的问题之一^[107]。

除了对于缺陷总数的预测，还有一些关于缺陷位置的预测研究。如，Rahman 等人比较了 SBF (Static Bug-Finding) 和 SDP (Statistical Defect Prediction)。前者是一种在编写代码的时候提示可能的缺陷及位置的工具，后者则通过建模预测代码在哪些地方可能存在缺陷。研究表明，两者的成本收益比相近，并且 SDP 可以提高 SBF 的表现^[148]。

3.1.6 资源估算

除了上述几种估算，还有资源估算^[11]。网络的发展在当今的产业中占有重要地位，

文献 [11] 深入研究了网络资源的估计，并给出了一个网络资源估计的系统评价，以确定当前网络资源估计的发展现状与可能存在的研究间的差距。他们首先确定当前的研究现状，在执行一个全面的文献检索后，选取了 84 篇经验研究，以便于调查网络资源估算的各个方面。研究结果表明，现今没有一个在特定估算场景下该使用哪些资源估算技术、怎样执行、怎么评价其效率的指导；精确度的结果也有很大的不同，并且依赖于许多因素。重心在开发工作量/成本估算的研究上，忽略了其他方面的资源估计，如质量和维护；规模的度量已经应用在很多地方，但是只在一个研究中被作为一个资源的指示器。他们的研究结果表明，在网络资源估算领域还有大量的工作可做。

3.1.7 国内研究现状

成本估算方面。西安交通大学的宋擒豹等^[167]关注于项目早期的特征子集选取和工作量估算，他们提出了一个使用灰色系统理论中的灰色关联分析的新颖方法，能够解决小样本的不确定性。接着，他们^[168]使用基于关联规则挖掘的方法去预测缺陷的关联性和缺陷修正工作量，并将该方法应用在超过 15 年的 200 个项目中。他们^[191]还提出了一种基于灰色模型 GM (1, 1) 和 Verhulst 的方法去预测软件物理时间的阶段工作量。在大型的现实世界的软件工程数据集上验证该方法，并且将其与线性回归方法和卡尔曼滤波方法相比，结果显示精确度至少分别提高了 28% 和 50%。

中国科学院软件研究所的李明树等^[101]全面回顾、分析了软件成本估算的各种代表性方法，也归纳、讨论了与成本估算强相关的软件规模度量问题，进一步研究了软件成本估算方法的评价标准，并指出软件成本估算方法下一步的主要发展趋势。他们还调研了中国工业界中影响开发工作量的因素，并在 140 个软件组织的 999 个项目中验证了 6 个因素（项目规模、团队规模、工期、开发类型、业务范围、编程语言）对开发工作量的影响。

缺陷预测方面。南京大学的周毓明等^[210]利用面向对象设计度量元预测高、低严重程度的缺陷。之后，文献 [212] 给出了一种新的指标来衡量名称相同的类在两个相邻版本中所发生的变化。他们^[109]还采用统计元分析方法在 102 个 Java 系统中去探究 62 个面向对象度量元预测易出缺陷类的能力^[109]。他们在研究^[186]中还利用方法调用异常来提高基于频谱的缺陷定位技术在面向对象程序中的有效性^[186]。南京大学的陈振宇等^[123-124]提出了一种基于聚类的策略来确定故障定位的准确性。

中国科学院软件研究所的张文、王青等提出了基于 K 近邻和社会网络分析的缺陷修复人推荐方法 Drex。接着，他们^[200]利用 LDA 主题模型针对不同程度开发人员对于修复缺陷的不同程度兴趣和专长进行建模，并在此基础上提出 DRETOM 方法以进行缺陷修复人推荐^[200]。

西安交通大学的宋擒豹等^[214]搜集了类级别的静态代码度量元和变化数据，并分析发现 80% 的代码行变化存在于 20% 的类中，然后他们使用分类技术来预测易改变的类并得到了良好的效果。

南京大学的周志华、清华大学的张洪宇等^[102]将主动学习、组合学习和半监督学习

的思想引入软件缺陷预测，针对可利用的软件缺陷历史数据稀少珍贵的问题，提出了基于抽样的主动式半监督缺陷预测方法。

清华大学的张洪宇等^[39]提出一种基于调用堆栈重复缺陷报告聚类的方法。他们的研究^[213]还参考缺陷报告的文本长度和缺陷报告之间的文本相似性，并提出了一种基于信息检索的缺陷定位方法 BugLocator。文献 [58] 还提出一种利用简单的用户反馈的交互式缺陷定位方法。进一步，文献 [65] 分析了一个缺陷报告是真正的缺陷还是一个过时的测试。文献 [197] 还涉及利用分段和堆栈跟踪分析来提高面向缺陷报告的缺陷定位的准确度。

南京大学的张赫等人^[108]采用系统评价的研究方法综合分析了目前 CRC 模型的相关研究^[108]。其结果表明，在基于 4 个基本 CRC 模型的 18 个具体的计算算子中，并没有一个相对确定的最好算子，并且，缺乏工业领域的应用也是 CRC 模型相关经验研究当前状态存在的问题之一。

3.2 软件需求、软件结构设计及度量

3.2.1 需求分析

软件需求分析是软件开发活动中的重要阶段，大量研究表明，修复该阶段发生的错误的代价远远高于其他阶段^[97]。目前，在需求工程领域的经验研究方面有如下代表性工作。

在需求捕获（Requirements elicitation）方面，Dieste 等学者采用系统评价方法搜索 SCOPUS、IEEE Xplore、ACM Digital Library 以及 Google 数据库，从找到的 564 篇论文中筛选出 26 篇含有需求捕获技术的经验研究。这些研究用来评估和评价 43 种需求捕获技术，并根据获得的结果提出了一整套指导需求捕获实践的原则和建议^[42]。有研究者提倡在需求分析阶段应充分发挥创造性以改善需求捕获的效果。Saha 等人搜索 IEEE Xplore、ACM、Compendex、Inspec、Springerlink、Science Direct 等不同的电子数据库，他们根据获得的结果得出创造性在需求工程中所起的作用^[152]。

鉴于用户往往是需求分析的主要参与和贡献者，Bano 等人研究用户参与需求分析活动的程度与项目成功之间的关系，他们从找到的 87 个经验研究中（1980 年 2012 年）发现仅有 13 个把重点放在用户参与需求分析活动中^[14]，所得结论是，在需求工程阶段有效地让用户参与可以减少其他阶段用户参与的必要性。此外，他们还提出了在需求分析中应该使用的能提高用户参与程度的推荐活动列表。

Svensson 等人发现需求工程研究者在如下 5 个方面进行了经验研究：需求捕获、依赖性、质量需求度量、成本估算及确定需求优先次序。此外，他们还研究了现有的质量需求管理方法的优点及局限性等问题^[177]。

Condori-Fernandez 等人采用系统评价方法，收集现阶段软件需求规约技术研究中的哪些方面已经进行了相关的经验研究、研究的背景是什么以及采用了什么研究方法等方

面的资料。他们分析了所找到的 46 篇相关论文，得出以下结论^[38]：大部分经验研究评估的焦点是软件需求规约的可理解性，实验是最常用的研究方法，评估的环境大部分是学术机构。

3.2.2 软件结构设计

软件体系结构是软件系统顶层的设计蓝图，是软件开发从需求到详细设计的中间产品。软件体系结构在软件开发过程的整个生命周期中，对软件质量、开发成本和进度的控制都起到关键作用^[15,120]。以软件体系结构为中心的软件开发（Architecture-centric software development）已经成为大型复杂软件系统^[15,159]的基石和准则，并得到一系列工业化软件开发过程（CMM、UP）、软件体系结构描述规范（ISO 42010：2011^[75]、ADL^[119]）、支撑工具（IBM rational software architect）和成熟框架平台（.NET、J2EE）的支持。

软件体系结构是软件开发的中间制品，具有一定程度的抽象性和多样性，一般由架构师来负责，设计过程本身难以完全自动化，因此软件体系结构领域的研究难点在于其方法和效果难以得到有效评价。经验软件工程强调方法在应用中的实际评价，这些方法包括调研、实验和案例研究等，因此它们也适用于软件体系结构领域的研究，并已得到广泛应用^[145,48]。

（1）体系结构的系统评价

软件体系结构的概念从 1992 年被提出以来^[141]，经历了 20 多年的不断发展，已经成为软件工程研究和实践中相对成熟的领域^[37]，在软件开发，尤其是大型软件开发中得到了广泛应用。系统评价（Systematic Review）作为经验软件工程中的一种主流文献综述方法，近年来被广泛应用于软件体系结构领域内特定研究主题的系统性研究和实践成果的提炼与呈现上。Williams 和 Carver 对软件系统演化过程中体系结构层面的变化进行了综述研究，将软件体系结构的变化提炼为 4 种：完善变化、修正变化、适宜性变化和预防性变化，并将它们用于软件体系结构层面的变化分析和处理中^[196]。Breivold 等则对软件系统演化中采用的软件体系结构方法进行了系统评价，提炼出了主流的基于软件体系结构的系统演化方法，并对所有方法进行了分类和综合分析^[26]。Oliveira 等对现有面向服务体系结构（SOA）的参考体系结构模型进行了提炼，并提出了一个综合面向服务体系结构的参考模型^[136]。Koziolek 从软件体系结构对系统的可持续性和生存性影响方面进行综述和分析，提出了一系列保持软件体系结构层面系统可生存的设计准则^[94]。Aleti 等对软件体系结构的优化方法（尤其是在大型复杂软件系统中，针对复杂的质量需求和目标，如何通过搜索技术得到最佳的软件体系结构设计方案）进行了系统综述，并基于综述结果为软件体系结构的优化指出了若干重要研究方向^[5]。Klein 和 van Vliet 对系统的系统（System-of-systems，SOS）领域的软件体系结构研究进行了综述，指出了该领域存在的若干挑战性问题，包括 SOS 的体系结构可扩展性和稳定性问题^[92]。李增扬等对知识方法应用于软件体系结构进行了综述，按照软件体系结构过程，针对知识方法对各个软件体系结构活动的支持进行综述，为知识方法在架构活动中的应用奠定了基础^[103]。Shahin 等对

可视化技术用于软件体系结构建模和描述进行了综述，提炼出了主流的软件体系结构可视化方法，为软件体系结构可视化实践提供了指南^[157]。Tofan 等对软件体系结构中的决策方法和实践进行了系统综述，提炼出各种方法的优势和不足，并为软件体系结构决策的下一步研究提出了可行的研究路线^[184]。

(2) 体系结构的工业调研

工业调研方法被用来反映当前工业界针对软件体系结构方法、概念的理解和应用状况。如 Malavolta 等对工业界目前使用和需要的软件体系结构描述语言的调研，发现了目前软件体系结构描述语言在工业应用中存在的主要问题，以及可能的应对方法^[117]。Tang 等调研了工业界如何理解、使用和归档软件体系结构设计原理，发现了该领域存在的主要挑战，同时提出了未来应解决的问题^[181]。Falessi 等比较了现有的用于制定软件体系结构决策的技术，并认为目前没有万能的方案，不同的技术适合解决不同的问题^[49]。Lehrig 和 Becker 对软件体系结构的受控实验进行了调研，总结出了相关的经验教训，并提出未来仍然需要实施更多的软件体系结构方面的受控实验^[98]。丁炜等调研了 2000 个开源软件项目，并发现仅 108 个项目有软件体系结构的文档，以及在这 108 个项目中模型、系统和任务是最常被记录的内容^[43]。Rost 等调研了 147 个工业界从业者在软件体系结构文档方面遇到的问题以及对于未来的期望，指出目前软件体系结构文档更新不及时，以至于和已有数据不一致的问题，同时还发现对目前软件体系结构文档的检索是一个重要的挑战^[151]。Tofan 等关注于软件体系结构的决策，通过调研 43 个工业界的架构师，从多个维度比较了这些架构师做出的决策，发现了该领域存在的一些问题^[183]。de Silva 和 Balasubramaniam 调研了防止、检测和恢复软件体系结构侵蚀的技术，并对这些技术进行了分类，总结出了各自的优势和劣势^[163]。van Heesch 和 Avgeriou 关注于制定软件体系结构决策的推理过程，通过对工业界 53 个架构师的调研，总结出了一些软件体系结构决策推理的最佳实践^[68]。

(3) 体系结构的受控实验

受控实验方法被用于评价软件体系结构方法在控制条件下对特定变量的应用效果。Javed 和 Zdun 发现了应用软件体系结构的追溯性可以极大地促进用户对软件体系结构层面的理解，同时发现个人经验对于理解软件体系结构并非决定性因素^[77]。Lytra 等研究了软件体系结构知识的复用对软件体系结构制定决策过程的影响，并发现复用决策模型可以显著提高新手架构师的工作效率和效果^[110]。Shahin 等研究了如何用软件体系结构的设计决策提高用户对软件体系结构的理解，并得到了若干结论，如发现了在软件体系结构文档中使用软件体系结构设计决策并不能降低完成软件体系结构设计任务的时间^[158]。De Graaf 等研究了基于本体的归档如何支持用户寻找所需的软件体系结构知识，并发现了对本体的构建如果是以对软件体系结构知识的良好理解为基础，则可以显著提高用户寻找所需软件体系结构知识的效率和效果^[60]。Haitzer 和 Zdun 关注于软件体系结构构件的图表对新手架构师理解软件设计的影响，并指出如果构件图表中的元素能直接链接到需要被理解的问题，则用软件体系结构构件的图表能增进新手架构师对软件设计和软件架构的理解^[63]。van Heesch 等验证了在恢复软件体系结构决策时应用软件体系结构模式

可以使软件体系结构决策恢复得更加有效这一假设，并得出通过应用软件体系结构模式可以显著提高决策质量这一结论，但没有证据支持用这一模式可以影响决策的数量^[69]。Ferrari 等识别了一些会被已有软件体系结构影响的需求特征，以及影响程度，同时还识别出了影响这些需求特征的软件体系结构的元素^[51]。Babar 等关注于在评价软件体系结构时对场景简介的开发，指出分布式的会议更加有利于场景简介的开发，但是这在很大程度上依赖于工具的支持^[12]。

(4) 体系结构的案例研究

案例研究方法被广泛应用于实际应用环境中对软件体系结构方法的评价。Moreno-Rivera 和 Navarro 对软件产品线体系结构应用于 Web 地理信息系统开发的方法进行了研究，发现该方法能够满足软件产品线体系结构设计中的关键性系统质量需求^[128]。López 等提出了一种基于网页的语义方法和工具（Toeska/Review），以比较和显示软件体系结构的原理，并在博物馆集成项目中（Contexta）使用^[111]。Buckley 等评估了专家架构师使用 Reflexion Modelling 方法时软件体系结构的一致性，并有针对性地对该方法进行了扩展^[28]。Schultis 等识别了 3 个协作模型以及一系列软件体系结构的挑战和问题，并对这些挑战进行了分类^[154]。Brunet 等分析了为期 5 年的 Eclipse 体系结构检查报告，识别出一些对开发者有用的规则，发现了代码是如何偏离期望的软件体系结构以及开发者是如何处理这些问题的^[27]。De Graaf 等提出了基于本体的软件体系结构归档方法，总结出 8 个可以影响本体成功构建的因素，并通过案例描述该方法是如何获得的，并对软件体系结构的知识需求建模^[59]。李增扬等提出了基于软件体系结构决策和变化场景的识别软件体系结构技术债（Architectural Technical Debt）的方法，并评估了该方法的有效性和可用性，发现该方法非常适合在工业案例中使用且对软件体系结构技术债的识别有效^[105]。李增扬等建议使用软件模块化度量标准来表示软件体系结构技术债，通过验证 ANMCC（Average Number of Modified Components per Commit）与模块化度量标准的联系，发现了两个模块化度量标准与 ANMCC 紧密关联，并可使用于软件体系结构技术债中^[104]。Díaz 等验证了一种用于敏捷构建和演化产品线体系结构的方法，并指出该方法能有效地促进从业者构建和演化产品线体系结构^[46]。Etemadi 等研究了元启发式优化方法是如何优化软件体系结构设计过程的，并通过案例发现用该方法可以找到满足所有质量属性和约束的有效解决方案^[47]。Mattsson 等报告了模型驱动开发在真实项目中的使用情况，发现了模型驱动开发在建模软件体系结构的设计规则方面的不足，并提出了相关建议^[118]。丁炜等通过分析开源项目中的开发邮件列表，发现邮件列表中的信息能将软件体系结构的变化反映到代码层面（但是大部分软件体系结构的变化都是预防性的），同时指出开源项目的软件体系结构在第一个稳定版本后趋于稳定^[44]。Nakagawa 等研究了软件体系结构如何直接与开源软件质量联系起来，并提出了使用软件体系结构重构来修复软件体系结构的方式^[131]。Feilkas 等通过研究软件体系结构知识的缺失问题来回答软件体系结构在多大程度上与代码一致；文档是否能够反应预想的软件体系结构等问题^[50]。

3.2.3 国内研究现状

武汉大学的梁鹏、李增扬等对知识方法应用于软件体系结构进行了综述，按照软件

体系结构过程，对知识方法针对各个软件体系结构活动的支持进行综述，为知识方法在架构活动中的应用奠定了基础^[103]。他们还提出了基于软件体系结构决策和变化场景的识别软件体系结构技术债的方法，并评估了该方法的有效性和可用性，发现该方法非常适合在工业案例中使用且对软件体系结构技术债的识别有效^[105]。接着，他们还建议使用软件模块化度量标准来表示软件体系结构技术债，通过验证 ANMCC 与模块化度量标准的联系，发现了两个模块化度量标准与 ANMCC 紧密关联，并可使用于软件体系结构技术债中^[104]。

武汉大学的丁炜、梁鹏等调研了 2000 个开源软件项目，发现仅 108 个项目有软件体系结构的文档，以及在这 108 个项目中模型、系统和任务是最常被记录的内容^[43]。他们还通过分析开源项目中的开发邮件列表，发现邮件列表中的信息能将软件体系结构的变化反映到代码层面（但是大部分软件体系结构的变化都是预防性的），同时指出开源项目的软件体系结构在第一个稳定版本后趋于稳定^[44]。

3.3 软件测试与质量保障

3.3.1 软件测试

近年来，经验方法在软件测试领域的研究和应用方面呈现越来越多的趋势，以 ICSE2015 大会为例，本次大会中在测试方面共有 16 篇论文，其中 12 篇均为采用实验分析的研究论文。软件测试的实验过程相对统一：选择实验程序、运行测试用例、收集测试结果、进行实验统计分析。

在软件测试实证分析中，影响结果的因素主要包括：实验程序（含测试用例）、评价标准（覆盖、变异、故障等）和统计方法（描述统计、推断统计）。我们以 ICSE2015 这 12 篇论文为例，按这 3 个因素分别讨论软件测试实证分析的现状。

（1）实验程序

在 12 个研究中，所有实验均采用开源实验程序。其中，6 篇论文采用 C 语言程序，5 篇论文采用 Java 语言程序。实验程序的来源主要是 SIR、GNU 及其他。50% 的实验都采用了 SIR^[164]上的程序作为实验对象。实验对象的规模则相差较大，从几十行到十万行代码不等。值得提醒的是，文献虽采用了较大规模的实验程序，但其测试主要基中于类这一级别，因而实际程序规模大幅减小。

实验程序的数目差异也较大。只有两篇论文^[40,198]选用了超过 30 个实验程序，满足基本统计数量要求。7 篇论文只采用了数量极少（10 个以内）的实验程序。现有论文在实验程序的选择上没有严格策略，基本采用前人常用的说辞。过少的实验程序数量和随意的选择策略，使得实验结果的外延风险增加，从而限制了实验结论的适用范围。测试用例是软件测试实验的必要组成部分。SIR 自带测试用例集，实验重现方便，因而广受研究人员欢迎。除此之外，采用工具产生测试数据也是一种常用手段，但这一手段通常局限在规模较小的白盒测试数据生成工具中，如 KLEE、EvoSuite、Randoop 等。人工生

成测试数据实验成本高，因而难以被广泛采用。综上所述，实验程序开放，但相应测试数据并未开放，这对研究论文的实验重现带来较大的困难。

(2) 评价标准

软件测试的最主要目标是发现失效行为，即俗称的 Bug。因而 Bug 检测能力成为软件测试方法的常用评价标准之一。12 篇论文中，有 3 篇采用真实 Bug 进行软件测试方法的评价^[140, 182, 198]。然而，我们不难发现，真实的 Bug 数量较少，因此难以提供有效的统计支撑。

在 SIR 的程序库中，大部分采用人工注入 Bug 作为实验程序。人工注入 Bug 的一种常用手法是变异，即通过机械式的语法改变进行 Bug 注入。变异分析存在成本昂贵和等价变异等困扰，使其难以在工业中应用，但在软件测试实证分析研究中，仍然不失为一种好的评价手段。其中文献 [40] 采用了变异分析方法。

在难以获取 Bug 信息的情况下，控制流覆盖和数据流覆盖是一种替代的软件测试评价手段。12 篇论文中，有 5 篇采用了覆盖信息作为评价标准^[53, 140, 194, 198, 178]。尽管基于覆盖的评价标准在学术界和工业界都得到了广泛采用。然而，近几年开始出现了批评的声音：为了覆盖而覆盖没有实际的 Bug 检测能力。

时清凯等提出一种距离熵的测试用例充分性准则，以期通过度量测试用例集的多样性来刻画其 Bug 检测能力^[161]。这一基于软件行为的度量方式为软件测试能力评价标准提供了新的思路。

(3) 统计方法

统计方法是实证分析的重要组成部分，通常分为描述统计与推断统计。描述统计通常采用统计图表、汇总数据等方式，用于直观展现实验结果的基本情况。均值和百分比是最常见的数据汇总方式，几乎出现在所有论文中。箱形图易于呈现实验结果的平均值和波动情况，散点图和折线图易于呈现实验结果的趋势走向和相关性，被广泛采用。

推断统计是利用样本数据来推断总体特征的统计方法。显著性检验是软件工程中常用的推断统计方法，包括 t - 检验和 U - 检验等。

3.3.2 质量保障

软件质量保障技术除了上述的软件测试之外，还包括静态程序分析、形式化规约、统计过程控制、软件评审等多种技术手段。在这些技术手段中，软件评审一直是经验软件工程研究的热点。软件评审一般是指应用人工阅读的方式静态地评价软件开发产物，以期尽量在项目早期发现并且消除缺陷，从而提升最终软件产品的质量。典型的评审对象包括各类开发文档和代码。评审的方式也有多样，例如，个人阅读、走查 (Walk Through)、正式评审 (Formal Inspection) 等。其中，正式评审自从 1976 年由 M. Fagan 引入软件开发以来，在工业界和学术界都引起了很多关注。Zhang 等人的调查表明，正式评审是软件工程实践中被研究和报告最多的软件工程实践之一^[207]。应用经验软件工程方法研究软件评审的实践主要包括如下一些具体的研究方向。

(1) 阅读技术

阅读是软件评审最基本的实践之一，相应地，各种阅读技术（Reading Techniques）也是软件评审相关研究中最活跃的研究课题之一^[153]。目前被识别的主要阅读技术包括：基于检查表的阅读（Checklist-based Reading, CBR）、基于场景的阅读（Scenario-Based Reading, SBR）、基于缺陷的阅读（Defect-based Reading, DBR）、基于视角的阅读（Perspective-based Reading, PBR）、基于用例的阅读（Usage-based Reading, UBR）、抽象驱动阅读 Abstraction-driven Technique (ADT)、任务驱动阅读（Task-driven Inspection, TDI）等。当然，还有最基本的随机阅读（Ad Hoc Reading, AHR）。

经验软件工程在这个方面的主要研究集中在针对不同的评审对象，比较不同阅读技术在缺陷发现率、缺陷发现成本等问题上的差异，从而试图找到一个具体场景下的较为合适的阅读技术。

例如，Caroline D. Rombach 等人通过软件工程实验对比了在 ER（实体关系）模型的设计文档评审中，用什么样的阅读技术可以更高效地发现 ER 模型中的缺陷。Thomas Thelin 等人则通过一个实验，比较了 UBR 和 CBR 阅读技术在需求评审中发现缺陷的时间效率和缺陷发现率上的差异。Maldonado 等人通过实验比较了 PBR 和 CBR 在发现缺陷上的差异^[81]。Oliver Laitenberger 等人通过准受控实验，比较了 PBR 和 CBR 对于代码评审在缺陷发现率上的差异^[135]。Alastair Dunsmore 等人通过实验，研究了针对面向对象代码的评审的 3 种阅读技术，即 CBR、UBR 和 ADT^[2]。

值得注意的是，Ciolkowski 等人的调查研究表明，在工业界，大部分评审者（60%）都缺乏必要的评审经验。在剩下 40% 的评审者当中，仅仅只有 12% 的人接受过正式的评审训练^[36]。因而，评审新手是一个普遍存在的现象。Rong 等人以软件企业中的评审新手为研究对象，对比 CBR 和 AHR 这两种不同的阅读技术，结论表明，就评审效果而言，两者没有显著差异，但是 CBR 可以提供更好的基础，有助于未来改进^[150]。

(2) 评审效果的影响因素

评审效果的影响因素相关的经验研究也非常丰富。一般而言，对于评审效果主要通过两个指标来衡量：即功效（一次评审发现的缺陷总数）和效率（单位时间内发现的缺陷数量）^[189]。目前经验方法研究所识别的主要因素包括：个人能力、评审速度、评审过程、准备阶段、材料总量、团队规模、人员培训、知识背景和经验、评审轮数以及材料类型等。通过这些研究，建立起对于影响评审效果的因素的相对丰富的理解。

例如，Biffl 等人通过实验并引入成本效益（Cost-Benefit）模型分析了阅读技术、评审团队规模、评审时间（速度）等因素对缺陷发现率和成本效益比的影响。研究表明，结合不同的阅读技术比单独采用最好的阅读技术拥有更高的效率；更好的评审过程比额外添加评审人员更有效；额外增加评审人员，并不能线性地提高评审效率。因此，为了实现较高的缺陷发现率所要付出的开销远大于平均水准的缺陷发现率对应的开销^[171]。

Albayrak 等人通过实验研究了个人的教育背景、经验、英语水平、公司等因素对于需求评审的影响。实验结果表明，在进行需求评审时，没有软件工程背景的评审者效率甚至更高；拥有更多的需求经验的评审者效率更高；评审者的英语水平和评审效率正相

关；评审者的公司背景对评审者评审效率有一定影响^[3]。

Sauer 等人基于行为学的理论进行了经验研究，研究指出，评审者个人的专业水平是影响评审效率的最重要的因素；提高评审效率有 3 种有效的方式，即，选择擅长发现缺陷的评审者、通过训练提高个人专业水平、选取适当的评审人数^[33]。

Kemerer 等人通过对个人软件过程（Personal Software Process）的数据研究，发现对于个人评审来说，在 200 LOC/hour 或稍低的评审速度下，评审效率最佳^[34]。

（3）支持工具的效果

工具支持主要是研究应用于软件评审过程中辅助评审者完成评审的软件系统和工具。该领域的研究起于 20 世纪 90 年代，也正是互联网应用迅速发展的时期。该领域的研究主要集中在两个方面：①研究针对不同的评审文档，工具支持的评审（Tool-based）和纸质评审（Paper-based）的差异；②为特定目的，提出新的工具，并进行验证。

Halling 等人通过 3 个独立实验（参与者、实验设计和执行均类似）比较了针对需求文档的工具支持评审和纸质评审这两种评审方式，得出以下结论：工具支持评审和纸质评审的缺陷发现率相近；工具支持评审可以降低所发现缺陷的重复率；此外，工具支持评审可以提高单位时间内发现的缺陷数^[61]。

Biffl 等人提出了 GRIP（Groupware-supported Inspection Process）框架并通过实验比较了在 GRIP 上进行评审和纸质评审的差异，证实了相较于纸质评审，GRIP 能够提高评审效率。GRIP 可以为评审团队提供一个框架并整合多种工具，它支持独立发现缺陷、评审会议以及评审管理等功能^[172]。

Perry 等人提出了 HyperCode（通过浏览器使用的评审工具），该工具可以帮助分散在各地的评审者进行非会议形式的评审。并通过实验证实了，相较于现存的纸质评审，HyperCode 的经济性明显更好；采用 HyperCode 可以缩短评审间的间隔；相较于现存的纸质评审，采用 HyperCode 可以发现更多的缺陷；此外，采用 HyperCode 之后，单个缺陷的描述更加清晰^[41]。

（4）其他

在评审子课题下，还有一些其他相关的经验软件工程研究课题，例如：Halling 等人通过一个经济模型研究了评审为软件开发带来的收益，其主要发现是，对于需求协商模型，评审是一种经济的验证技术^[62]。Ciolkowski 等人通过两个调查研究了软件评审实践在工业界中的应用状态，研究表明，在软件开发项目中，评审的目的从早期发现缺陷到使团队更好交流，各不相同。此外，目前评审中的阅读方法很多，但是，非系统化的阅读技术应用得最为广泛^[36]。

Denger 等人通过实验比较代码评审和功能测试这两种缺陷发现技术对软件开发中的重用组件的不同类型缺陷的发现率，根据实验结果发现，这两种技术对于变异缺陷（Variant-specific Defects）的识别率都很低^[35]。

3.3.3 国内研究现状

正式评审自从 1976 年由 M. Fagan 引入软件开发以来，在工业界和学术界都引起了很

多关注。南京大学的张赫等人的调查表明，正式评审是软件工程实践中被研究和报告最多的软件工程实践之一^[208]。南京大学的荣国平等人以软件企业中的评审新手为研究对象，对比 CBR（基于检查表的阅读）和 AHR（随机阅读）两种不同阅读技术，相关结论表明，就评审效果而言，两者没有显著差异，但是 CBR 可以提供更好的基础，有助于未来改进^[150]。南京大学的陈振宇等提出一种距离熵的测试用例充分性准则，以期通过度量测试用例集的多样性来刻画其 Bug 检测能力^[161]。这一基于软件行为的度量方式为软件测试能力评价标准提供了新的思路。

3.4 软件开发与过程改进

软件过程是软件工程领域的一个重要组成成分，对于软件开发的进度和质量都有着重要的影响。在过去的几十年里，软件过程中产生了大量的方法和技术。具有代表性的是基于计划的软件过程方法，与之对应的是敏捷开发方法。敏捷开发由于其不确定性，很多实践者往往从个人的经验和观点出发进行争论，缺乏严谨的科学基础。也有很多学者和实践者并不认同敏捷软件开发，他们认为没有足够的科学证据支持敏捷开发所宣称的益处。经验软件工程方法通过经验研究来确定敏捷软件开发的有效性和局限性，同时也展示当前敏捷软件开发在工业界的使用情况，以供实践者参考，为其科学的决策和改进提供了有效的技术与方法。

3.4.1 敏捷方法的经验研究

伴随着敏捷软件开发本身的发展，对于敏捷方法的研究，早期集中在 XP 方法，后来扩展到 Scrum 和 Kanban 方法。

Dyba 和 Dingsoyr 在 2008 年总结了当时在敏捷软件开发中进行的经验研究情况。在 2005 年以前，共有 36 篇关于敏捷软件开发的经验研究论文，其中有 3 篇是二手研究，33 篇研究中有 25 篇是关于 XP 方法的。Dyba 认为敏捷软件开发领域的经验研究可以分为①敏捷方法的引入和采纳；②人和社会性因素；③客户和开发者感受；④比较研究。作者认为这些研究确定了一些敏捷软件开发的益处和局限性，但是这些经验研究提供的证据强度都比较弱，不足以提供非常有效的建议供工业界使用。

Scrum 方法是当前应用最为广泛的敏捷方法。文献 [222] 讨论了 Scrum 在全球软件开发（GSD）中的情况，认为①对于在 GSD 的软件项目中使用 Scrum 的原因和动机缺少研究；②当前 GSD 中项目团队使用 Scrum 实践的情况和本地使用有所不同；③GSD 团队使用 Scrum 时通常会面对交流、合作等问题；④Scrum 实践需要扩展和修改来支持 GSD。

对于敏捷软件开发，很多人都认为其不适合于①大型项目和组织；②地理位置分散的团队；③对于软件质量和安全要求很高的项目。一些研究通过经验方法研究并介绍了在大型组织中进行敏捷软件开发的有效途径。

全球软件开发对于当前世界的全球化进程和软件外包产业非常重要，但是进行全球化开发会面临语言文化差异、信任问题、跨时空交流沟通等问题。一些研究通过经验研

究的方法讨论了如何在分布式团队环境下进行敏捷软件开发。

敏捷软件开发提倡更少的度量，研究发现使用度量的行为主要集中在以下领域：sprint 计划、进度跟踪、软件质量度量、修复软件过程问题、激励开发者。此外发现，尽管敏捷团队使用很多敏捷文献中建议的度量，但是他们同样也采用很多团队定制的度量。另外，最有影响的度量是开发速度和工作量估算。文章发现，度量在敏捷开发中的使用和传统软件开发相似：项目和 sprint 需要被计划和追踪；质量需要被度量；过程中的问题需要被识别和解决。

3.4.2 敏捷实践的经验研究

经验研究方法已被研究界广泛应用于研究敏捷实践。经验研究对于敏捷实践的研究主要集中于质量检查和结对编程（Pair Programming, PP）中。经验研究分析了大量的工业界项目，结果表明，在敏捷实践中：时间盒、计划会议、学习循环、演进和分级需求、每日讨论、产品愿景等实践被广泛采用，其他采用较多的有客户参与、质量检查。同时，结果还表明，敏捷实践已经在自动化、公司管理、咨询、金融、政府、互联网、医疗、媒体、通信等多行业的软件开发中被广泛采用。

软件质量是软件产品中非常关键的问题之一。测试驱动开发（Test-Driven Development, TDD）由 Kent Beck 在 2000 年提出，它是敏捷软件开发的一项基本实践，同时也是极限编程中非常重要的一部分。测试驱动开发方法要求开发者在编写具体功能前先编写测试代码，再编写功能代码。这种方法驱动着代码的设计与功能的实现、测试以及驱动代码的重构。虽然 TDD 的问世时间并不长，却在学术界引起了广泛研究。长期以来，在 TDD 经验研究中，实验和案例分析这两种经验研究方法被广泛采用。

经验研究对 TDD 的研究目标主要集中在 TDD 的有效性和适用性方面。有效性研究的主要目标为：对软件质量的影响、对生产率的影响、对软件设计的影响、对软件维护的影响。

敏捷需求工程被用来定义敏捷的计划、执行和解释需求工程活动，此外，以协作为中心的敏捷方法中的需求工程活动所引起的问题仍然没有被广泛认知。

3.4.3 软件过程改进

软件过程改进是一个实践和研究并重的课题。各种改进模型的提出、裁剪和应用形成了这个领域大部分的研究工作，具体包括以下几个典型的方向。

软件过程理论兴起于 20 世纪 80 年代，经过不断的发展和完善，已经形成一套比较成熟的理论。一些典型的模型也得到了广泛的应用。经验软件工程在过程模型方面的研究主要是各类软件过程改进（Software Process Improvement, SPI）模型在不同规模的公司及组织中的适用性，提高软件产品的质量或者加速软件开发过程。

Staples 等人的案例研究表明，CMMI 模型作为一个软件过程改进框架，不仅可以让一些大型组织通过达到高级别的成熟度获得高生产率和产品质量，同时 CMMI 也可以很好地适用于中小型软件企业^[170]。Niazi 等人在 11 个企业访谈结果的基础上，设计并且验证了一个软件过程改进的模型^[134]。将 CMM/CMMI 和其他技术融合也是此类研究的关注

重点，例如，Park 等人研究了 Six Sigma 在 PSP/TSP 过程中的应用，结合实例证实，Six Sigma 可以使软件工程师更好地分析 PSP 数据，并系统地改进过程性能^[139]。

在软件过程管理中要考虑很多关键问题，以对过程改进展开有效的建模和评估。然而，庞大的元素数量和多样性使得这种考虑面临很多挑战。García 等人提出并且验证了一种 FMESP 框架，即过程和测量软件建模的集成管理框架^[54]。每一个过程改进项目都必须优先选择并集中于过程的特定方面。此外，进行每一个此类选择都需要沟通、激励所有的利益相关者（管理以及参与或受影响的各方）。Birkholzer 等人提出了一个概念框架和工具集，结合实例分析，证实该框架可以用于目标驱动的搜索策略，以及特定场景的仿真分析^[20]。Jedlitschka 提出了一个基于经验改进的集成软件过程改进框架，并在工业界和学术界分别开展了案例研究和受控实验研究^[78]。

某些时候，全面实施 CMMI 框架看起来负载过重，对此，Sivashankar 等人提出并且验证了一个框架来更好地应用 CMMI 模型，这有助于实现 SPI 与一些标准的实施^[162]。Harjumaa 等人介绍了一组模式，以指导小公司软件检验过程的改进，初步的实验结果表明该模式提供了一种可行的快速启动 SPI 活动的方法^[66]。

Bayona 等人提出涉及流程改进的关键成功因素包括：承诺、精准的业务战略和目标、培训、交流、资源、技能、员工参与、改进管理、定义过程、软件过程改进流程的监测、变化管理、文化、政策、角色和责任、工具和指导的规定，以在技术上指导软件过程改进的方法。此外，在此模式下，由于员工参与过程改进活动，所以可以最大限度地减少改进阻力，保持员工的积极性。当然，过程改进计划还需要资源、人的技能、能力和知识来执行活动^[17]。Georg Kalus 等人以过程裁剪的参考标准为依据，应用系统评价方法研究，得出了类似的结论^[56]。Rong 等人进一步提出这些裁剪的标准必须是在一定上下文之下才有意义，因而，在应用经验软件工程方法进行相关研究的时候，必须定义具体的上下文。Michael 等人则提出定义、选择和验证措施可能比 SPI 评估更有意义等^[125]。

Beijun Shen 等人以一个小软件公司为背景，提出了一种软件过程改进的实施方式，实施措施主要包括过程建模、过程自动化，以及过程测量。Serrano 等人同样描述了在一个小的软件开发公司的过程改进经验中，使用团队软件过程（TSP）自主地实施软件过程改进，而在组织层，应用软件能力成熟度模型（CMM）以及 IDEAL 模型作为组织改进模型^[156]。Sulayman 等人应用系统评价方法研究软件过程改进在中小型网络公司中的应用。这些中小型公司往往具备类似特点，例如，严格的预算约束、严格的期限和短期战略要求等^[176]。

应用支持工具能够很好地支持、管理并规范软件过程，因此应用支持工具研究在软件工程改进相关的研究中也是数量较多的一类。

García、Du 等人报告了一些案例研究，目的是探讨和评价支持工具对过程改进的作用^[55,45]。

3. 4. 4 国内研究现状

上海交通大学的沈备军等以一个小软件公司为背景，提出了一种软件过程改进的实

施方式，这类方式结合了灵活和控制的考虑，并且不妨碍一个小公司的创新性质，实施措施主要包括过程建模、过程自动化，以及过程测量^[223]。

中国科学院软件研究所的王青等基于软件项目管理和过程改进工具 Qone 在使用中产生的数据，通过进行后续的文件调研和访谈，详细分析了软件过程管理工具的使用情况以及带来的启示，例如该工具的作用随着任务的类型变化而变化；对细粒度的任务更容易进行预测和控制^[45]。

3.5 开源与分布开发

在 20 世纪末，开源软件系统取得了巨大成功，已经成为软件技术创新和产业发展的主要模式之一，展现出社会、经济、组织与管理、技术、实践等方面的重要属性。这为寻求高效的软件开发方法提供了一种用户创新驱动、成本低、质量高的新思路。为探索高效的软件生产方法，有很多的报告致力于研究开源实践。同时，开源软件的数据开放性也为经验研究在此领域的发展提供了条件。目前，针对开源软件开发的经验研究主要存在两种类型：一类是基于对各类开发数据的分析与挖掘，发现模式；另一类是提出新的理论或模型，并使用开源经验数据进行验证。

开源已经成为软件技术创新和产业发展的主要模式，展现出社会、经济、组织与管理、技术、实践等方面的重要属性。从经济学角度，研究者关注的核心问题之一是：为什么会有大量的个体自愿且无偿地参与到开源软件开发中。从组织学的角度，研究者则关注：为什么一个看似松散无序的群体能够生产出高质量的软件制品。从技术的角度，研究者则关注：开源软件开发背后蕴含/基于的方法和技术是什么，使用了哪些支持工具。从实践角度，研究者则关注：商业型组织如何有效参与到开源软件开发中。

开源实践的一个重要特点是全球分布式开发和大众化协同。分布在世界各地的开发者协同发布一个可用的高质量软件时，会面临大规模的通信、协调和合作的挑战，这体现了社会化开发的复杂性。如何利用软件社区中的海量软件数据和丰富的软件知识，如何度量对程序员的技能、成长途径、他们跟环境的交互、环境对他们的影响，如何进一步理解群体构造的机理和演化规则，如何解决社会化开发的可知与可控难题，这些在目前都受到了广泛关注。

3.5.1 开源软件的开发和演化

Scacchi^[154] 基于对已有经验研究的回顾，研究开源软件是如何开发和演化的，主要包括：为什么人们会参与开源软件开发，资源和能力支持的开发过程，在项目中如何进行协同和控制，项目内部的信息沟通和社会网络，开源软件作为多项目的软件生态系统，开源软件作为社会运动，并且发现开源软件研究中未来的机会。

软件演化在开放的软件开发中面临着越来越多的挑战，涉众的多元化、开发的社区化、环境的开放性都给软件的演化带来了新的问题，Scacchi^[154] 发现开源软件项目在进行演化时，不是依赖于用户提出的形式化的功能性需求文档，而是依赖于在线制品数据，

例如需求请求、论坛中的信息、聊天脚本等。Pagano 等研究 AppStore 中的用户评论，发现这些评论内容包括软件需求和用户体验，例如已有功能的不足、需求请求、功能的使用场景等，但是这些评论的质量却差别很大。他们指出如果有工具或方法来系统地过滤和聚类这些用户评论，可以节省项目开发人员的工作量，从而有效地指导软件演化。为了节省花费在收集和理解用户评论上的时间，Guzman 等^[224]提出了一种自动化方法，以帮助项目开发人员过滤、组织和分析用户评论。该方法首先用自然语言理解的方法识别评论中的细粒度的需求信息，然后抽取出有这些需求的用户的偏好并且进行打分，最后用主题模型技术将细粒度的需求组织为有意义的高层需求集合。类似地，Chen 等^[225]开发了 AR-Miner，以过滤掉噪声评论、抽取信息量大的评论、对评论进行聚类以及优先级排序，从而指导软件演化。

Zimmermann 等^[215]利用版本控制系统数据来寻找软件代码中的变化模式（Change Pattern），进而对软件维护中代码的修改提供自动的启发式提示。依靠版本变动历史信息构建事务，并在事务基础上进行挖掘以构造代码维护的规则集合，进而利用规则集合来对代码维护进行指导。Howison 等^[71]研究了使用 SourceForge 数据挖掘可能遇到的问题和易犯错误的项目，作者指出使用 SourceForge 中开源项目数据时所需要注意的一些事项，包括在获取数据、分析数据时需要考虑哪些问题等，为利用开源项目数据提供了有益的指导。Fischer 等^[52]以 BSD 产品族（FreeBSD、Mac OS 等）为实验对象，突破了以单个项目历史信息为研究对象的局限，研究对一个产品族历史信息的挖掘和利用，结果表明，用他们的方法可以发现不同项目之间的代码依赖关系。Huang 等^[72]研究了如何对版本控制系统信息进行分析整理，为新进入开源项目开发的项目外围人员提供最初的学习指导。

重构是开源软件演化的重要方式，Ratzinger 等^[149]通过对开源软件的分析发现开发的生命周期属性与重构有关，因此提出了一种通过挖掘软件的演化信息来预测重构的方法。Taneja 等^[180]提出了用 refacLib 工具来自动地探测软件库中的重构，通过对重构前后两个版本的代码的语义分析和启发式的方法探测重构。最后通过对 5 个开源软件库的重构探测验证了方法的有效性。Liu 等^[107]提出了一种重构的调度方法，并利用开源项目的数据来验证方法的有效性。Palomba 等^[137]提出了一种基于代码变更历史信息来探测 code smell 的方法，并通过 8 个开源项目进行验证。Bavota 等^[16]提出了基于关联的主题模型的重构推荐方法。

为了研究重构是否能提高软件的质量，Alshayeb 等^[6]通过经验研究得到重构并不一定能够提高质量属性的结论。Murphy-Hill 等^[130]分析了重构工具的使用和代码中的重构类型，及重构的分布情况。结果表明重构在项目中发生得非常频繁，大部分重构都是零散地分布在其他开发活动中，但是重构工具很少被使用。Rachatasumrit 等^[146]分析了重构对回归测试的影响，结果显示 22% 的重构在回归测试中被测试了，并且有一半失败的回归测试涉及了重构。Murgia 等^[129]分析得到了代码之间的 fan-in、fan-out 与重构行为的相关关系。

3.5.2 开放社区的任务分配和调度

在开源软件开发过程中，群体智慧的来源是开源社区的开发者，开发者在开源项目

中的角色、能力和开发者之间的合作网络影响着软件开发的效率和质量。为了分析网络中人员的重要性，Lim 等^[99]通过建立涉众之间的网络，并分析网络度量指标来为涉众进行优先级排序。Yu 等^[204]提出了一种模型来描述开源软件中开发人员的交流，并利用数据挖掘技术来分析开发人员在组织中的角色。该方法通过对两个开源项目的分析验证了方法的有效性。Soh 等^[166]通过对 Mylyn 中记录的操作交互的历史信息，来分析开发人员在进行任务维护时的工作量。Blincoe 等^[21]分析任务间的依赖关系与合作关系，并且发现 17% 的任务需要合作。

为了探索开源软件社区中是否存在子社团，以及这种结构是否有利于用户更好地理解自组织的软件社团的工作，Bird 等^[23]通过对 Apache 的开发邮件和软件资产库中的活动分析，重新组织志愿者的网络。然后利用复杂网络理论评价该网络并观察其随时间的变化。研究结果显示，在 Apache 中确实存在一些子社团，并且项目的年龄与这些子社团的精密程度有关联。Xu 等^[201]通过对 SourceForge 里的开发人员网络进行拓扑分析和演化统计分析，来了解开源软件的现象。研究结果显示，开源软件中的开发人员网络是一个无标度网络。

在开源社区任务分配和推荐方面，Kagdi 等^[84]提出了一种为软件变更推荐开发人员的方法。该方法根据历史变更代码提交记录，分析变更经验、经历，以及开发人员的贡献，并以此推荐相关的开发人员。在并行的软件开发中，尤其是开源社区中分布式异地的群体开发中，尽管有事先的沟通和协商，但冲突的变更依然会存在。Kasi 等^[86]分析了开源项目中同时变更代码的冲突，并提出了一种最小化冲突的调度方法，最后用开源项目进行验证。

软件缺陷的修复是一项困难、成本高昂且漫长的过程，快速准确地将缺陷报告分派给合适的开发人员可以大大降低缺陷的修复时间。

Xie^[200]等提出了一种基于主题模型的软件缺陷修复人员推荐方法（DRETOM），该方法利用开发人员参与缺陷修复的历史活动数据，对缺陷报告进行主题建模，并进一步计算开发人员在这些主题上的兴趣度和能力经验值及其参与修复新缺陷报告的概率，最终推荐排名较高的开发人员作为潜在的修复人员。

Xuan 等^[202]提出基于开发者排序的缺陷报告分派方法，该方法首先根据开发者在缺陷跟踪系统中共同评论缺陷报告的信息构建开发者社会网络，然后根据开发者的贡献对开发者进行排序。首先使用传统的机器学习缺陷报告分派方法给缺陷报告推荐 N 个修复人，然后根据开发者在开发者社会网络中的顺序重新对这 N 个修复人进行排序。Naguid 等提出了一种基于开发人员历史活动集合的缺陷报告分派方法，他们将开发人员在社区中不同模块里的活动分为“review”、“assign” 和 “resolve” 3 种类型。首先计算出某个开发者在某个模块上这 3 种活动的不同概率。当某个模块上出现新的缺陷报告时，用该方法结合开发者历史活动，推荐一组“resolve” 概率最大的开发人员来修复缺陷报告。Wang 等^[192]提出一种基于异构网络模型的缺陷报告分派方法（DevNet），用以表征缺陷数据库中开发者之间的多种不同类型的协作关系，并在此基础上结合历史缺陷报告修复信息，将新缺陷报告准确地分派给合适的缺陷修复人员。

3.5.3 开源社区的人员成长

开发者加入开源社区一般都会遵循社会化过程，也称为“洋葱模型”。新加入的开发者首先通过邮件列表的讨论和缺陷跟踪等活动，锻炼自己的能力，在社区中获得声誉，然后逐渐成为核心成员，从而可以直接修改代码和进行设计决策。

项目参与者成为核心代码贡献者是他的技术贡献和社会交互的双重作用结果。Gharehyazie 等^[57]认为，对于技术贡献的判断，需要提取不同来源的补丁提交数据，这个较难获取，而社会交互数据较易得到，因此，他们意在研究只基于社会交互数据，能够多大程度地对开发者状态进行建模。通过基于 6 个 Apache 开源项目的研究，发现用社会交互数据，特别是双向的交流数目，可以显著地预测开发者是否能够成为核心代码贡献者。该研究还描述了在成为核心代码贡献者的临近时间点上，开发者的社会交互模式。

当新手加入开源项目时，他们进行适当的培训，从而使其了解该项目的技术问题和组织结构。Canfora 等^[29]提出 Yoda 模型，基于邮件列表和版本控制系统，为开源软件项目的新手识别和推荐导师。Yoda 模型在 5 个开源软件项目上被验证，该研究还通过调研这些项目的开发者来了解指导活动是否真的在这些项目中存在。Steinmacher 等^[173]研究阻碍新手进行第一次贡献的社会障碍，具体通过系统文献调研，开源项目开发者的问卷调研，以及半结构化的访谈，提出了一个包含 58 个障碍（其中有 13 个社会障碍）的概念模型。新手面临的社会障碍包括：接待方面（例如没有收到回复）、交流方面（进行无用的评论、羞怯等）、方向指导方面（例如如何找到导师）、文化差异方面（例如发送的消息看起来显得粗鲁）等。

3.5.4 开源社区的协作模式及影响

在分布式开发环境中，协作和交互对于完成复杂交错的任务是至关重要的。协作和交互的模式及效率会对开发者的生产率以及软件产品的质量产生影响。

Pinzger^[144]等基于开源软件的版本库，用社会网络分析方法，研究开源软件的沟通和协作，并开发了 STNA-Cockpit 工具，该工具不仅可以提供元模型来表示沟通和协作，并且能够用图形可视化技术来发现元模型的实体，由此得到在某段时间内项目的动态变化。

Bettenburg 和 Hassan^[18]研究软件开发中的社会交互活动（Social Interactions）对软件质量的影响。该研究基于问题跟踪系统的讨论信息，分为三个部分，一是基于讨论信息得到的软件开发的社会结构（例如开发者在社会网络中的中心度）对软件质量的影响；二是讨论的内容和属性（例如所提交的补丁包含的文件数目）对软件质量的影响；三是讨论活动反映出的团体的工作流（例如缺陷报告中的工作流活动）对软件质量的影响。

项目团队的结构对于项目的表现有何影响？项目团队内部和外部开发者之间的沟通是如何影响项目的最终成败的？为了回答这些问题，Hossain 等^[70]建立缺陷管理过程的协作网络，并且研究其中的结构属性对项目表现的影响。结果表明，很多网络指标（例如密度、中心性等）和缺陷管理活动的质量、时间存在相关性。Bhattacharya 等^[19]建立缺陷修复过程中的开发者协作关系网和代码提交过程中的开发者协作关系网，并且研究

这些网络的网络指标和缺陷严重程度、模块的维护工作量，以及缺陷数目之间的关系。

“社会 - 技术一致性”（Socio-Technical Congruence）是通过技术的依赖关系和项目成员的交互程度来衡量团队的协作情况。Kwan 等^[96]研究团队的协作情况与软件构建活动的关系，结果发现对于持续构建，高的“社会 - 技术一致性”会带来更大概率的成功；而对于集成构建，高的“社会 - 技术一致性”反而会导致成功的概率变小。Cataldo 和 Herbsleb^[31]研究“社会 - 技术一致性”对软件质量和开发生产率的影响。结果发现，协作需要和实际的协作活动之间的差异会大大提高缺陷发生的概率；并且高的“社会 - 技术一致性”会带来开发者生产率的提升。他们还发现，技术依赖和实际的协作活动之间的一致性不管对于成熟的软件项目，还是对于刚起步的开发环境，都是很重要的。

虽然开发者协作不畅会带来诸如集成失败、重复工作、进度延期、软件缺陷等问题，但是开发者有些时候还是对协作的必要性认识不足。Blincoe 等^[22]提出一种方法，能够即时高效地识别和推荐协作。该方法通过捕获集成开发环境中的开发者行为，运用任务的属性和机器学习方法进行协作的精准推荐，从而预防信息过载。该方法在 Mylyn 项目中通过用户访谈和定量的分析进行验证。

3.5.5 基于群体智慧的众包开发

众包属于新兴商业模式，李未院士将众包软件开发称为“群体软件开发”，并提出“群体软件工程”的基本内容，其理论核心是：从封闭走向开放、从精英走向大众、从机械工程到社会工程。其开发原则为“使用者即设计者，使用者即开发者，使用者即维护者”，该开发原则目前主要在 App Store 应用程序中得以体现。目前国内对于众包领域的研究主要集中在众包概念推广、众包应用模式、众包对社会变革可能带来的影响、众包与外包的联系与区别等理论层面，少部分经验研究主要集中在以 AMT（Amazon Mechanical Turk）为代表的微任务（例如图像识别、创意征集、Logo 设计等）领域。而对复杂任务，特别是众包软件开发的研究较少，已有的研究也大多处于理论层面，缺乏经验及数据支持。

Yang 等^[203]基于微任务众包平台 Taskcn 分析了众包接包方的行为模式，并归纳其学习过程与行动策略，为众包任务设计提供了参考建议。Stolee 等^[174]针对经验软件研究中数据难以获取的问题，提出通过众包平台征集接包方以收集实验数据，但收集到的数据质量需要进一步被验证。

目前针对众包软件开发方面的研究主要是基于 TopCoder 众包软件开发平台的：Archak^[10]将经济学中的“廉价磋商”概念引入众包平台，并探讨荣誉机制与众包软件开发质量之间的关系，该研究表明高等级用户之间存在着“廉价磋商”行为，基于此可自动优化众包社区中的资源配置，并且高等级用户更倾向于提交高质量的软件制品。Mao 等^[116]基于众包软件开发历史数据从项目类型、输入质量、输入复杂度、前阶段决策 4 个方面提出共 16 个成本因子，进一步使用多种统计及机器学习方法建立经验定价模型并进行了验证，实验结果表明，所提出的经验定价模型具有较高的准确性，并可以通过回归因子显著性分析抽取出简单易用的定价规则。Li 等^[226]分析了影响众包软件开发质量的

因素，并从众包平台以及任务的角度提出了 23 个质量因子，同时进一步通过回归分析识别出平台任务的平均质量、近期任务数量、文档长度等关键质量因子，并为改善众包软件质量提供了建议。

对于众包群体推荐方面的研究目前也尚处于起步阶段，已知现有的研究是基于 AMT 平台的微任务众包，Ambati 等^[7] 基于群体兴趣与技能采用隐式建模技术进行群体推荐，并通过实验说明其推荐行为可以提高众包任务的完成质量。Yuen 等^[206] 基于群体过去的任务偏好（例如任务的类别）和表现（例如针对某一类别任务收到的酬金与花费的时间等）提出了一种群体推荐方法，指出该方法可以激励众包群体持续参与众包工作并提升工作质量。

3.5.6 国内研究现状

针对当前研究对于各种 bad smells 的检测和重构都是独立进行的现状，北京大学的邵维忠等^[107] 提出了一种检测和消除序列方法，该方法针对不同类型的 bad smell，能够简化其检测和消除过程。该方法首先基于示例说明采用检测和消除序列的必要性，并且对于经常发生的 bad smell，推荐了检测和消除序列。

对于缺陷修复人推荐问题，中国科学院软件研究所的王青等^[200] 提出了一种基于主题模型的软件缺陷修复人员推荐方法（DRETOM），该方法利用开发人员参与缺陷修复的历史活动数据对缺陷报告进行主题建模，然后计算开发人员在这些主题上的兴趣度和能力经验值，最终推荐潜在的修复人员。另外，他们^[192] 还提出一种基于异构网络模型的缺陷报告分派方法（DevNet），用以表征缺陷数据库中开发者之间多种不同类型的协作关系，并在此基础上结合历史缺陷报告修复信息，对缺陷修复人进行推荐。

大连理工大学的江贺等^[202] 提出基于开发者排序的缺陷报告分派方法，该方法首先根据开发者在缺陷跟踪系统中共同评论缺陷报告的信息构建开发者社会网络，然后根据开发者的贡献对开发者进行排序。在推荐时，首先使用传统的机器学习缺陷报告分派方法给缺陷报告推荐 N 个修复人，然后根据开发者在开发者社会网络中的顺序重新对 N 个修复人进行排序。

众包方面，中国科学院软件研究所的王青等^[116] 基于众包软件开发历史数据从项目类型、输入质量、输入复杂度、前阶段决策 4 个方面提出共 16 个成本因子，进一步使用多种统计及机器学习方法建立经验定价模型并进行了验证，实验结果表明，所提出的经验定价模型具有较高的准确性，并可以通过回归因子显著性分析抽取出简单易用的定价规则。另外，他们^[226] 还分析了影响众包软件开发质量的因素，并从众包平台以及任务的角度提出了 23 个质量因子，进一步通过回归分析识别出平台任务的平均质量、近期任务数量、文档长度等关键质量因子并为改善众包软件质量提供了建议。

4 国内外研究工作比较和展望

国内在经验软件工程方面的研究起步于 21 世纪初。一些研究团队，如中国科学院软

件研究所、北京大学、南京大学、清华大学、武汉大学、上海交通大学、复旦大学、北京航空航天大学、西安交通大学、东南大学等开展了卓有成效的工作。CCF 的软件工程专委会于 2013 年成立经验软件工程学组，本次报告的撰写也主要来自该学组的成员。

其中，中科院软件所在成本估算、缺陷预测、需求演化、众包分析以及软件过程和质量改进方面开展了较多国际先进的工作，相关工作在 ICSE、RE、ESEM、ICSM、ICSSP 等会议以及 ASE、ESE、IST 等期刊都有报告；北京大学在基于代码版本库的软件缺陷分析、基于开源代码的微过程挖掘与分析等方面取得了国际领先的研究成果，并发表在 ICSE、FSE、ASE 等软件工程顶级会议上；南京大学在过程改进、质量保障、测试以及缺陷预测等方面的工作上取得很好的进展，相关工作结果在 ICSE、FSE、ESEM、ICSSP 等高水平会议上有报告；复旦大学在基于大规模源码的代码演化分析、代码查找与推荐等方面取得了显著的进展，并将相关结果发表在 ICSE、FSE 等软件工程顶级会议上；武汉大学在需求获取、软件体系结构、缺陷预测以及开源社区分析等方面的工作中也取得了非常好的成果，相关工作结果在 ICSE、FSE、ICSSP 等会议以及 ASE、JSS、IST、TSC 等期刊上都有发表。

CCF 软件工程专委会成立的经验软件工程学组在推动经验软件工程学科的发展，特别是在大学的课程教育方面起到了积极的推动力作用。中国科学院软件研究所、北京大学、南京大学还是国际软件工程研究联盟 ISERN (International Software Engineering Research Network) 的正式会员。正是该联盟组织积极推动了经验软件工程学科的发展，本领域最著名的学者（如 Barry Boehm、Victor Basili、Dieter Rombach 等）都是该联盟的创始成员。

此外，中国的软件产业发展很快，百度、华为等一批企业对经验软件工程的实践需求也开始突显，为我国经验软件工程领域的发展提供了良好的环境和机遇。以上研究团队也都在不同程度上建立了与我国软件企业，特别是龙头企业的合作关系，一些研究成果已经在企业中获得成功应用。

5 结束语

经验软件工程越来越明显地成为软件工程领域主要的研究方法和应用领域，这一结论在近年来的研究论文中得到广泛认可。我国在经验软件工程方面的起步较晚，但机遇和发展空间良好。一些非常好的研究工作也赶上国际先进水平。

总体而言，软件工程也是一门实验科学的认识在最近越来越被大家所认同，特别是在大数据风起云涌的今天，软件的开放性使得在开发、使用、维护和演化各个层面的不确定性日渐突出。而现代质量管理理论认为对数据和信息的逻辑分析和直觉判断是有效决策的基础。经验软件工程从观察、假设、分析到推导结论的方法，为解决开放、不确定性、大数据环境下的软件质量问题提供了有效的方法和途径。但软件工程有其特殊性，主要表现在：

- 稀疏性。大部分的软件工程数据在其关注的属性方面都具有典型的二八性质，譬

如缺陷数据，20%的模块包含80%的缺陷。数据的不均衡性使很多统计方法的应用受到限制。

- 独特性。软件是人类智能的产物，没有任何项目是可以复制的，软件工程的数据也受制于项目、产品、组织的特征，存在严重的数据漂移和语义不统一的问题。
- 可解释性。软件工程数据分析的动机是理解现象产生的原因，支持合理的决策，并开展科学的过程和产品的改进。因此我们进行的实验和分析必须是可解释的，而不能仅限于统计意义。

这些问题影响经验软件工程研究是否有效可用的关键。此外，对于研究方法，还要在以下几个方面共同努力：

- 经验研究数据。需要更为广泛和丰富的Benchmark实验程序。目前PROMISE、NASA、Ecilips、ISBSG等提供了一些公共数据集，但面对开放环境，这些还不够。大多数的经验研究都需要“爬取”数据。建立一些公共的开放实验库，对经验软件工程的研究将非常有益。
- 经验分析大多数依然停留在简单的描述统计分析上。这是必要的但还不够，我们应该采用更加系统的统计或机器学习方法，完成精确的分析，使其成为一个严谨的工程学科。经验软件工程学组翻译了一本在国际上被很多大学广泛使用的教材《Experimentation in Software Engineering》，希望在大学相关的课程中就开始训练学生们严谨的经验软件工程方法。
- 经验分析的原型工具和实验数据开发程度远远不够。譬如软件测试需要动态执行程序收集实验数据，这给实验重现带来极大挑战。我们应该提倡研究人员在发表高水平论文的同时，开放原型工具和实验数据，开放的科研成果有助于加强科研合作交流、科研人才培养以及增强科研能力。

此外，随着技术的发展、大数据时代的到来，我们可以从实践中获得海量的数据，但在大数据的环境下，再人工地理解每个项目的应用领域和最佳实践以获得对研究结果的解释将不太可能。要想从这些海量数据中把分散的隐藏的信息挖掘出来，就需要使用数据挖掘相关的技术和算法。与机器学习、社会化网络分析、数据挖掘、自然语言处理、信息检索等学科门技术的结合，可以帮助我们以更丰富的手段和方法获得有价值的信息，并建立有效的理论与预测模型。如何结合软件过程数据的自身特质，选用合适的、在已有的基础上改进的，甚至突破传统的数据挖掘和分析方法，将是非常重要和有意义的。

致谢

本报告在撰写的过程中得到软件工程专委会，尤其是经验软件工程学组许多老师的大力支持，特别是周毓明、彭蓉、王俊杰、王丹丹、杨晨等多位老师和同学，他们都为本文的撰写做出了很多贡献，在这里我们一并表示衷心的感谢！

参考文献

- [1] Akiyama F. An Example of Software System Debugging[C/OL]. IFIP Congress (1). 1971 , 71 : 353-359.
- [2] Dunsmore A, Roper M, Wood M. Practical code inspection techniques for object- oriented systems: an experimental comparison[J/OL]. IEEE software , 2003 (4) : 21-29. http://www2. dc. ufscar. br/~elis_hernandes/phdproposal/systematicmappings/Q%205.1%20papers/Dunsmore-Practical%20code%20inspection%20techniques%20for%20object-oriented%20systems%20An%20experimental%20comparison-2003.pdf.
- [3] Albayrak Ö, Carver J C. Investigation of individual factors impacting the effectiveness of requirements inspections: a replicated experiment[J/OL]. Empirical Software Engineering , 2014 , 19(1) : 241-266. <https://www.cs.gmu.edu/~offutt/classes/763/papers/Albayrak-reqs-EMSE2014.pdf>.
- [4] Albrecht A J, Gaffney Jr J E. Software function, source lines of code, and development effort prediction: a software science validation[J/OL]. Software Engineering, IEEE Transactions on , 1983 (6) : 639-648.
- [5] Aleti A, Buhnova B, Grunske L, et al. Software architecture optimization methods: A systematic literature review[J/OL]. Software Engineering, IEEE Transactions on , 2013 , 39(5) : 658-683.
- [6] Alshayeb M. Empirical investigation of refactoring effect on software quality [J/OL]. Information and software technology , 2009 , 51 (9) : 1319- 1326. <http://alshayeb.com/publications/J04-Empirical%20Investigation%20of%20Refactoring%20Effect%20on%20Software%20Quality.pdf>.
- [7] Ambati V, Vogel S, Carbonell J G. Towards Task Recommendation in Micro- Task Markets [C/OL]. Human computation. 2011 : 1-4. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.210.5588&rep=rep1&type=pdf>.
- [8] Amou M, Salehie M, Tahvildari L. Temporal software change prediction using neural networks[J/OL]. International Journal of Software Engineering and Knowledge Engineering , 2009 , 19(07) : 995- 1014. <http://www.stargroup.uwaterloo.ca/~msalehie/papers/IJSEKE09.pdf>.
- [9] Anvik J, Hiew L, Murphy G C. Who should fix this bug? [C/OL]. Proceedings of the 28th international conference on Software engineering. ACM , 2006 : 361-370. <http://www.st.cs.uni-sb.de/edu/empirical-se/2006/PDFs/p361.pdf>.
- [10] Archak N. Money, glory and cheap talk: analyzing strategic behavior of contestants in simultaneous crowdsourcing contests on TopCoder. com[C/OL]. Proceedings of the 19th international conference on World wide web. ACM , 2010 : 21-30. <http://www.ra.ethz.ch/cdstore/www2010/www/p21.pdf>.
- [11] Azhar D, Mendes E, Riddle P. A systematic review of web resource estimation[C/OL]. Proceedings of the 8th International Conference on Predictive Models in Software Engineering. ACM , 2012 : 49-58. http://www.researchgate.net/profile/Damir_Azhar/publication/236900605_A_Systematic_Review_of_Web_Resource_Estimation/links/0c96051a2a50f49cb0000000.pdf.
- [12] Babar M A, Kitchenham B, Jeffery R. Comparing distributed and face- to- face meetings for software architecture evaluation: A controlled experiment[J/OL]. Empirical Software Engineering , 2008 , 13(1) : 39-62. <http://eprints.lancs.ac.uk/64500/>.
- [13] Banker R D, Kauffman R J, Kumar R. An empirical test of object-based output measurement metrics in a

- computer aided software engineering (CASE) environment [J/OL]. Information Systems Working Papers Series , Vol , 1991. <http://dl.acm.org/citation.cfm?id=155421>.
- [14] Bano M, Zowghi D. Users' involvement in requirements engineering and system success [C/OL]. Empirical Requirements Engineering (EmpiRE), 2013 IEEE Third International Workshop on. IEEE, 2013 : 24-31. <https://opus.lib.uts.edu.au/research/handle/10453/27523>.
- [15] Bass L. Software architecture in practice, 3rd Edition [M/OL]. Pearson Education India, 2007. http://www.researchgate.net/profile/Rick_Kazman/publication/224001127_Software_Architecture_in_Practice/links/02bfe510fef5da3230000000.pdf.
- [16] Bavota G, Oliveto R, Gethers M, et al. Methodbook: Recommending move method refactorings via relational topic models [J/OL]. Software Engineering, IEEE Transactions on, 2014, 40(7) : 671-694. <https://www.infona.pl/resource/bwmeta1.element.ieee-art-000006684534>.
- [17] Bayona S, Calvo-Manzano J A, San Feliu T. Review of Critical Success Factors Related to People in Software Process Improvement [M/OL]. Systems, Software and Services Process Improvement. Springer Berlin Heidelberg, 2013 : 179- 189. https://books.google.com/books?hl=zh-CN&lr=&id=nF25BQAAQBAJ&oi=fnd&pg=PA179&ots=ER7_MijG9m&sig=wezEdxI8xlJ9dRRKTllemiv02zEg.
- [18] Bettenburg N, Hassan A E. Studying the impact of social structures on software quality [C/OL]. Program Comprehension (ICPC), 2010 IEEE 18th International Conference on. IEEE, 2010 : 124-133. <http://www.computer.org/csdl/proceedings/icpc/2010/4113/00/4113a124-abs.html>.
- [19] Bhattacharya P, Iliofoiu M, Neamtiu I, et al. Graph-based analysis and prediction for software evolution [C/OL]. Proceedings of the 34th International Conference on Software Engineering. IEEE Press, 2012 : 419-429. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6227173&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6227173.
- [20] Birkhölzer T, Dickmann C, Vaupel J. A Framework for Systematic Evaluation of Process Improvement Priorities [C/OL]. Software Engineering and Advanced Applications (SEAA), 2011 37th EUROMICRO Conference on. IEEE, 2011 : 294- 301. <http://www.computer.org/csdl/proceedings/seaa/2011/4488/00/4488a294-abs.html>.
- [21] Blincoe K, Valetto G, Damian D. Do all task dependencies require coordination? the role of task properties in identifying critical coordination needs in software projects [C/OL]. Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering. ACM, 2013 : 213- 223. <https://www.cs.drexel.edu/~kac358/publications/FSE2013-Blincoe.pdf>.
- [22] Blincoe K, Valetto G, Damian D. Facilitating Coordination between Software Developers: A Study and Techniques for Timely and Efficient Recommendations [J/OL]. IEEE Trans. Softw. Eng. 99, (May 2015), 1- 16. http://www.researchgate.net/profile/Giuseppe_Valetto/publication/276206809_Facilitating_Coordination_between_Software_Developers_A_Study_and_Techniques_for_Timely_and_Efficient_Recommendations/links/555281cb08ae980ca606b023.pdf.
- [23] Bird C. Community structure in oss projects [R/OL]. Technical report, University of California, Davis, 2006. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.4758&rep=rep1&type=pdf>.
- [24] Boehm B W, Madachy R, Steele B. Software cost estimation with Cocomo II with Cdrom [M/OL]. Prentice Hall PTR, 2000. <http://www.citeulike.org/group/19011/article/13312192>.
- [25] Boehm B, Abts C, Chulani S. Software development cost estimation approaches—A survey [J/OL]. Annals of software engineering, 2000, 10(1-4) : 177-205. <http://www.yamaghani.com/Files/01122012114937.pdf>.

- [26] Breivold H P, Crnkovic I, Larsson M. A systematic review of software architecture evolution research [J/OL]. *Information and Software Technology*, 2012, 54 (1) : 16- 40. <http://romisatriawahono.net/lecture/rm/survey/software%20engineering/Software%20Architecture/Breivold%20-%20Software%20architecture%20evolution%20research%20-%202012.pdf>.
- [27] Brunet J, Murphy G C, Serey D, et al. Five years of software architecture checking: A case study of Eclipse[J/OL]. 2014. <http://www.dcc.ufmg.br/~mtov/mes/seminarios/9.pdf>.
- [28] Buckley J, Ali N, English M, et al. Real-Time Reflexion Modelling in architecture reconciliation: A multi case study[J/OL]. *Information and Software Technology*, 2015, 61 : 107- 123. <http://eprints.brighton.ac.uk/13500/>.
- [29] Canfora G, Di Penta M, Oliveto R, et al. Who is going to mentor newcomers in open source projects? [C/OL]. Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering. ACM, 2012: 44. <http://www.gerardocanfora.net/preprints/whoisgoingtomentornewcomersinopensourceprojectsbygcanforamidipentarolivetoandspanichella/FSE%202012.pdf>.
- [30] Cao L, Mohan K, Xu P, et al. How extreme does extreme programming have to be? Adapting XP practices to large-scale projects[C/OL]. System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference on. IEEE, 2004: 10 pp. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.117.5437&rep=rep1&type=pdf>.
- [31] Cataldo M, Herbsleb J D. Coordination breakdowns and their impact on development productivity and software failures[J/OL]. *Software Engineering, IEEE Transactions on*, 2013, 39 (3) : 343-360. <http://herbsleb.org/web-pubs/pdfs/Cataldo-Coordination-2013.pdf>.
- [32] Chaumun M A, Kabaili H, Keller R K, et al. A change impact model for changeability assessment in object-oriented software systems[C/OL]. *Software Maintenance and Reengineering*, 1999. Proceedings of the Third European Conference on. IEEE, 1999: 130- 138. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.448.1130&rep=rep1&type=pdf>.
- [33] Sauer C, Jeffery D R, Land L, et al. The effectiveness of software development technical reviews: A behaviorally motivated program of research[J/OL]. *Software Engineering, IEEE Transactions on*, 2000, 26 (1) : 1-14. <http://www.computer.org/csdl/trans/ts/2000/01/e0001.pdf>.
- [34] Kemerer C F, Pault M C. The impact of design and code reviews on software quality: An empirical study based on PSP data [J/OL]. *Software Engineering, IEEE Transactions on*, 2009, 35 (4) : 534- 550. <http://search.proquest.com/docview/195582447?pq-origsite=gscholar>.
- [35] Denger C, Kolb R. Testing and inspecting reusable product line components: first empirical results [C/OL]. Proceedings of the 2006 ACM/IEEE international symposium on Empirical software engineering. ACM, 2006: 184-193. http://www2.de.ufscar.br/~elis_hernandes/phdproposal/systematicmappings/Q%205.1%20papers/Denger-Testing%20and%20Inspecting%20Reusable%20Product%20Line-2006.pdf.
- [36] Ciolkowski M, Laitenberger O, Biffl S. Software reviews: The state of the practice[J/OL]. *IEEE software*, 2003 (6) : 46-51. <http://search.proquest.com/docview/215842934?pq-origsite=gscholar>.
- [37] Shaw M, Clements P. The golden age of software architecture[J/OL]. *Software, IEEE*, 2006, 23 (2) : 31-39. <http://cat.inist.fr/?aModele=afficheN&cpsidt=17558587>.
- [38] Condori-Fernandez N, Daneva M, Sikkel K, et al. A systematic mapping study on empirical evaluation of software requirements specifications techniques[C/OL]. Proceedings of the 2009 3rd International Symposium

- on Empirical Software Engineering and Measurement. IEEE Computer Society, 2009: 502-505. http://www.researchgate.net/profile/Maya_Daneva/publication/221494786_A_systematic_mapping_study_on_empirical_evaluation_of_software_requirements_specifications_techniques/links/0deec517a94b33f918000000.pdf.
- [39] Dang Y, Wu R, Zhang H, et al. ReBucket: a method for clustering duplicate crash reports based on call stack similarity [C/OL]. Proceedings of the 34th International Conference on Software Engineering. IEEE Press, 2012: 1084-1093. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.464.2676&rep=rep1&type=pdf>.
- [40] Denaro G, Margara A, Pezze M, et al. Dynamic Data Flow Testing of Object Oriented Systems [C/OL]. 37th International Conference on Software Engineering, ICSE. 2015, 15. <http://www.utdallas.edu/~lxz144130/cs6301-readings/test-generation-denaro-icse15.pdf>.
- [41] Perry D E, Porter A, Wade M W, et al. Reducing inspection interval in large-scale software development [J/OL]. Software Engineering, IEEE Transactions on, 2002, 28(7): 695-705. <http://search.proquest.com/docview/195569606?pq-origsite=gscholar>.
- [42] Dieste O, Juristo N. Systematic review and aggregation of empirical studies on elicitation techniques [J/OL]. Software Engineering, IEEE Transactions on, 2011, 37(2): 283-304. <https://www.infona.pl/resource/bwmeta1.element.ieee-art-000005416730>.
- [43] Ding W, Liang P, Tang A, et al. How do open source communities document software architecture: An exploratory survey [C/OL]. Engineering of Complex Computer Systems (ICECCS), 2014 19th International Conference on. IEEE, 2014: 136-145. <http://www.computer.org/csdl/proceedings/iceccs/2014/5482/00/5482a136-abs.html>.
- [44] Ding W, Liang P, Tang A, et al. Causes of Architecture Changes: An Empirical Study through the Communication in OSS Mailing Lists [OL]. <http://www.cs.rug.nl/search/uploads/Publications/ding2015cac.pdf>.
- [45] Du J, Yang Y, Lin Z, et al. A Case Study on Usage of a Software Process Management Tool in China [C/OL]. Software Engineering Conference (APSEC), 2010 17th Asia Pacific. IEEE, 2010: 443-452. <http://yadda.icm.edu.pl/yadda/element/bwmeta1.element.ieee-000005693221>.
- [46] Díaz J, Pérez J, Garbajosa J. Agile product-line architecting in practice: A case study in smart grids [J/OL]. Information and Software Technology, 2014, 56(7): 727-748. <https://www.infona.pl/resource/bwmeta1.element.elsevier-2981a10d-ec2c-3fbf-92f9-111474e9d914>.
- [47] Etemaadi R, Lind K, Heldal R, et al. Quality-driven optimization of system architecture: Industrial case study on an automotive sub-system [J/OL]. Journal of Systems and Software, 2013, 86(10): 2559-2573. <https://www.infona.pl/resource/bwmeta1.element.elsevier-c22b306f-c0f8-3080-9bd8-b6a9174f8a50>.
- [48] Falessi D, Babar M A, Cantone G, et al. Applying empirical software engineering to software architecture: challenges and lessons learned [J/OL]. Empirical Software Engineering, 2010, 15(3): 250-276. http://www.researchgate.net/profile/Philippe_Kruchten/publication/225678298_Applying_empirical_software_engineering_to_software_architecture_challenges_and_lessons_learned/links/0912f51086a16b9907000000.pdf.
- [49] Falessi D, Cantone G, Kazman R, et al. Decision-making techniques for software architecture design: A comparative survey [J/OL]. ACM Computing Surveys (CSUR), 2011, 43(4): 33. <http://cat.inist.fr/?aModele=afficheN&cpsidt=25254432>.
- [50] Feilkas M, Ratiu D, Jürgens E. The loss of architectural knowledge during system evolution: An industrial

- case study [C/OL]. Program Comprehension , 2009. ICPC' 09. IEEE 17th International Conference on. IEEE , 2009 : 188-197. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5090042&tag=1.
- [51] Ferrari R , Miller J A , Madhvji N H . A controlled experiment to assess the impact of system architectures on new system requirements [J/OL]. Requirements Engineering , 2010 , 15 (2) : 215-233. <http://link.springer.com/article/10.1007/s00766-010-0099-3>.
- [52] Fischer M , Oberleitner J , Ratzinger J , et al . Mining evolution data of a product family [M/OL]. ACM , 2005. <http://dl.acm.org/citation.cfm?id=1083145>.
- [53] Gao Z , Liang Y , Cohen M B , et al . Making System User Interactive Tests Repeatable : When and What Should we Control ? [J/OL]. <http://cse.unl.edu/~myra/papers/gao-icse15.pdf>.
- [54] García F , Piattini M , Ruiz F , et al . FMESP : Framework for the modeling and evaluation of software processes [J/OL]. Journal of Systems Architecture , 2006 , 52 (11) : 627-639. <http://www.sciencedirect.com/science/article/pii/S1383762106000658>.
- [55] Pacheco C . A Web- based Tool for Automatizing the Software Process Improvement Initiatives in Small Software Enterprises [J/OL]. Latin America Transactions , IEEE (Revista IEEE America Latina) , 2010 , 8 (6) : 685-694. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5688096.
- [56] Kalus G , Kuhrmann M . Criteria for software process tailoring : a systematic review [C/OL]. Proceedings of the 2013 International Conference on Software and System Process . ACM , 2013 : 171-180. <http://dl.acm.org/citation.cfm?id=2486078>.
- [57] Gharehyazie M , Posnett D , Filkov V . Social activities rival patch submission for prediction of developer initiation in oss projects [C/OL]. Software Maintenance (ICSM) , 2013 29th IEEE International Conference on. IEEE , 2013 : 340-349. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6676905.
- [58] Gong L , Lo D , Jiang L , et al . Interactive fault localization leveraging simple user feedback [C/OL]. Software Maintenance (ICSM) , 2012 28th IEEE International Conference on. IEEE , 2012 : 67-76. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6405255.
- [59] De Graaf K A , Liang P , Tang A , et al . An exploratory study on ontology engineering for software architecture documentation [J] . Computers in Industry , 2014 , 65 (7) : 1053-1064.
- [60] de Graaf K A , Liang P , Tang A , et al . Supporting architecture documentation : a comparison of two ontologies for knowledge retrieval [C] . Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering . ACM , 2015 : 3.
- [61] Halling M , Biffl S , Grünbacher P . An experiment family to investigate the defect detection effect of tool-support for requirements inspection [C] . Software Metrics Symposium , 2003. Proceedings. Ninth International. IEEE , 2003 : 278-285.
- [62] Halling M , Biffl S , Grünbacher P . An economic approach for improving requirements negotiation models with inspection [J] . Requirements Engineering , 2003 , 8 (4) : 236-247.
- [63] Haitzer T , Zdun U . Controlled experiment on the supportive effect of architectural component diagrams for design understanding of novice architects [M] . Software Architecture . Springer Berlin Heidelberg , 2013 : 54-71.
- [64] Han A R , Jeon S U , Bae D H , et al . Measuring behavioral dependency for improving change- proneness prediction in UML-based design models [J] . Journal of Systems and Software , 2010 , 83 (2) : 222-234.
- [65] Hao D , Lan T , Zhang H , et al . Is this a bug or an obsolete test ? [M] . ECOOP 2013- Object- Oriented Programming . Springer Berlin Heidelberg , 2013 : 602-628.

- [66] Harjumaa L, Tervonen I, Vuorio P. Improving software inspection process with patterns [C]. Quality Software , 2004. QSIC 2004. Proceedings. Fourth International Conference on. IEEE , 2004: 118-125.
- [67] He M, Zhang H, Yang Y, et al. Understanding the influential factors to development effort in Chinese software industry[M]. Product-Focused Software Process Improvement. Springer Berlin Heidelberg , 2010: 306-320.
- [68] van Heesch U, Avgeriou P. Mature architecting- a survey about the reasoning process of professional architects[C]. Software Architecture (WICSA), 2011 9th Working IEEE/IFIP Conference on. IEEE , 2011: 260-269.
- [69] van Heesch U, Avgeriou P, Zdun U, et al. The supportive effect of patterns in architecture decision recovery—A controlled experiment[J]. Science of Computer Programming , 2012, 77(5) : 551-576.
- [70] Hossain L, Zhu D. Social networks and coordination performance of distributed software development teams [J]. The Journal of High Technology Management Research , 2009, 20(1) : 52-61.
- [71] Howison J, Crowston K. The perils and pitfalls of mining SourceForge[C]. Proceedings of the International Workshop on Mining Software Repositories (MSR 2004). 2004: 7-11.
- [72] Huang S K, Liu K. Mining version histories to verify the learning process of legitimate peripheral participants[J]. ACM SIGSOFT Software Engineering Notes , 2005, 30(4) : 1-5.
- [73] IFPUG F. International Function Point Users Group (IFPUG) Function Point Counting Practices Manual [J/OL]. 2000. <http://www.ifpug.org/itips-vtips/>.
- [74] Isong B, Obeten E. A SYSTEMATIC REVIEW OF THE EMPIRICAL VALIDATION OF OBJECT-ORIENTED METRICS TOWARDS FAULT- PRONENESS PREDICTION [J]. International Journal of Software Engineering and Knowledge Engineering , 2013 , 23(10) : 1513-1540.
- [75] Systems and software engineering: architecture description[S]. ISO , 2011.
- [76] Jabangwe R, Börstler J, Šmite D, et al. Empirical evidence on the link between object-oriented measures and external quality attributes: A systematic literature review[J]. Empirical Software Engineering , 2013 , 20(3) : 640-693.
- [77] Javed M, Zdun U. The supportive effect of traceability links in architecture-level software understanding: Two controlled experiments[C]. Software Architecture (WICSA), 2014 IEEE/IFIP Conference on. IEEE , 2014: 215-224.
- [78] Jedlitschka A, Pfahl D. Experience- based model- driven improvement management with combined data sources from industry and academia [C]. Empirical Software Engineering , 2003. ISESE 2003. Proceedings. 2003 International Symposium on. IEEE , 2003: 154-161.
- [79] Jeffery R, Ruhe M, Wieczorek I. A comparative study of two software development cost modeling techniques using multi-organizational and company-specific data[J]. Information and software technology , 2000, 42(14) : 1009-1016.
- [80] 贾浩. 软件缺陷分类预测的对比研究[D]. 中国科学院研究生院 , 2011.
- [81] Maldonado J C, Carver J, Shull F, et al. Perspective-based reading: a replicated experiment focused on individual reviewer effectiveness[J]. Empirical Software Engineering , 2006, 11(1) : 119-142.
- [82] Jørgensen M, Indahl U, Sjøberg D. Software effort estimation by analogy and “ regression toward the mean ”[J]. Journal of Systems and Software , 2003, 68(3) : 253-262.
- [83] Jørgensen M. A review of studies on expert estimation of software development effort[J]. Journal of Systems and Software , 2004, 70(1) : 37-60.

- [84] Kagdi H, Hammad M, Maletic J. Who can help me with this source code change? [C]. Software Maintenance , 2008. ICSM 2008. IEEE International Conference on. IEEE, 2008: 157-166.
- [85] Karner G. Resource estimation for objectory projects[J]. Objective Systems SF AB, 1993, 17.
- [86] Kasi B K, Sarma A. Cassandra: Proactive conflict minimization through optimized task scheduling[C]. Software Engineering (ICSE) , 2013 35th International Conference on. IEEE, 2013: 732-741.
- [87] Kitchenham B A, Taylor N R. Software project development cost estimation[J]. Journal of Systems and Software , 1985 , 5(4) : 267-278.
- [88] Kitchenham B. A procedure for analyzing unbalanced datasets [J]. Software Engineering, IEEE Transactions on , 1998 , 24(4) : 278-301.
- [89] Kitchenham B, Mendes E, Travassos G H. A systematic review of cross-vs. within-company cost estimation studies[C]. Proceedings of the 10th international conference on Evaluation and Assessment in Software Engineering. British Computer Society , 2006: 81-90.
- [90] Kitchenham B, Mendes E, Travassos G H. Cross versus within- company cost estimation studies: A systematic review[J]. Software Engineering, IEEE Transactions on , 2007 , 33(5) : 316-329.
- [91] Kim S, Zhang H, Wu R, et al. Dealing with noise in defect prediction [C]. Software Engineering (ICSE) , 2011 33rd International Conference on. IEEE, 2011: 481-490.
- [92] Klein J, van Vliet H. A systematic review of system-of- systems architecture research[C]. Proceedings of the 9th international ACM Sigsoft conference on Quality of software architectures. ACM, 2013: 13-22.
- [93] Koru A G, Liu H. Identifying and characterizing change- prone classes in two large- scale open- source products[J]. Journal of Systems and Software , 2007 , 80(1) : 63-73.
- [94] Koziolek H. Sustainability evaluation of software architectures: a systematic review[C]. Proceedings of the joint ACM SIGSOFT conference-- QoSA and ACM SIGSOFT symposium- - ISARCS on Quality of software architectures-- QoSA and architecting critical systems-- ISARCS. ACM, 2011: 3-12.
- [95] 库燕.一种基于用例的成本估算方法及工具[D].中国科学院研究生院, 2011.
- [96] Kwan I, Schröter A, Damian D. Does socio-technical congruence have an effect on software build success? a study of coordination in a software project[J]. Software Engineering, IEEE Transactions on , 2011 , 37 (3) : 307-324.
- [97] Lawrence B, Wiegers K, Ebert C. The top risk of requirements engineering[J]. Software , IEEE , 2001 , 18(6) : 62-63.
- [98] Lehrig S, Becker S. Software Architecture Design Assistants Need Controlled Efficiency Experiments: Lessons Learned from a Survey[C]. Proceedings of the 1st International Workshop on Future of Software Architecture Design Assistants. ACM , 2015: 19-24.
- [99] Lim S L, Quercia D, Finkelstein A. StakeNet: using social networks to analyse the stakeholders of large- scale software projects [C]. Proceedings of the 32nd ACM/IEEE International Conference on Software Engineering- Volume 1. ACM , 2010: 295-304.
- [100] Li W, Henry S. Object-oriented metrics that predict maintainability[J]. Journal of systems and software , 1993 , 23(2) : 111-122.
- [101] 李明树, 何梅, 杨达, 等. 软件成本估算方法及应用 [J]. 软件学报 , 2007 , 18(4) : 775-795.
- [102] Li M, Zhang H, Wu R, et al. Sample-based software defect prediction with active and semi-supervised learning[J]. Automated Software Engineering , 2012 , 19(2) : 201-230.
- [103] Li Z, Liang P, Avgeriou P. Application of knowledge- based approaches in software architecture: A

- systematic mapping study [J]. *Information and Software Technology*, 2013, 55(5) : 777-794.
- [104] Li Z, Liang P, Avgeriou P, et al. An empirical investigation of modularity metrics for indicating architectural technical debt [C]. Proceedings of the 10th international ACM Sigsoft conference on Quality of software architectures. ACM, 2014; 119-128.
- [105] Li Z, Liang P, Avgeriou P. Architectural Technical Debt Identification based on Architecture Decisions and Change Scenarios [C]. Proceedings of the 12th Working IEEE/IFIP Conference on Software Architecture (WICSA), At Montréal, Canada, pp.211-219.
- [106] Liu Q, Mintram R C. Preliminary data analysis methods in software estimation [J]. *Software Quality Journal*, 2005, 13(1) : 91-115.
- [107] Liu H, Ma Z, Shao W, et al. Schedule of bad smell detection and resolution: A new way to save effort [J]. *Software Engineering, IEEE Transactions on*, 2012, 38(1) : 220-235.
- [108] Liu G, Rong G, Zhang H, et al. The adoption of capture-recapture in software engineering: a systematic literature review [C]. Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering. ACM, 2015; 15.
- [109] Lu H, Zhou Y, Xu B, et al. The ability of object-oriented metrics to predict change-proneness: a meta-analysis [J]. *Empirical software engineering*, 2012, 17(3) : 200-242.
- [110] Lytra I, Gaubatz P, Zdun U. Two controlled experiments on model-based architectural decision making [J]. *Information and Software Technology*, 2015, 63 : 58-75.
- [111] López C, Inostroza P, Cysneiros L M, et al. Visualization and comparison of architecture rationale with semantic web technologies [J]. *Journal of Systems and Software*, 2009, 82(8) : 1198-1210.
- [112] Ma Y, Luo G, Zeng X, et al. Transfer learning for cross-company software defect prediction [J]. *Information and Software Technology*, 2012, 54(3) : 248-256.
- [113] MacDonell S G, Shepperd M J. Comparing Local and Global Software Effort Estimation Models-- Reflections on a Systematic Review [C]. *Empirical Software Engineering and Measurement*, 2007. ESEM 2007. First International Symposium on. IEEE, 2007; 401-409.
- [114] Mair C, Shepperd M. The consistency of empirical comparisons of regression and analogy-based software project cost prediction [C]. *Empirical Software Engineering*, 2005. 2005 International Symposium on. IEEE, 2005; 10 pp.
- [115] Malhotra R. A systematic review of machine learning techniques for software fault prediction [J]. *Applied Soft Computing*, 2015, 27 : 504-518.
- [116] Mao K, Yang Y, Li M, et al. Pricing crowdsourcing-based software development tasks [C]. Proceedings of the 2013 international conference on Software engineering. IEEE Press, 2013; 1205-1208.
- [117] Malavolta I, Lago P, Muccini H, et al. What industry needs from architectural languages: A survey [J]. *Software Engineering, IEEE Transactions on*, 2013, 39(6) : 869-891.
- [118] Mattsson A, Lundell B, Lings B, et al. Linking model-driven development and software architecture: a case study [J]. *Software Engineering, IEEE Transactions on*, 2009, 35(1) : 83-93.
- [119] Medvidovic N, Taylor R N. A classification and comparison framework for software architecture description languages [J]. *Software Engineering, IEEE Transactions on*, 2000, 26(1) : 70-93.
- [120] 梅宏, 申峻嵘. 软件体系结构研究进展 [J]. *软件学报*, 2006, 17(6) : 1257-1275.
- [121] Mendes E, Di Martino S, Ferrucci F, et al. Effort estimation: how valuable is it for a web company to use a cross-company data set, compared to using its own single-company data set? [C]. Proceedings of

- the 16th international conference on World Wide Web. ACM, 2007: 963-972.
- [122] Mendes E, Di Martino S, Ferrucci F, et al. Cross-company vs. single-company web effort models using the Tukutuku database: An extended study [J]. Journal of Systems and Software, 2008, 81 (5): 673-690.
- [123] Miao Y, Chen Z, Li S, et al. Identifying Coincidental Correctness for Fault Localization by Clustering Test Cases[C]. SEKE. 2012: 267-272.
- [124] Miao Y, Chen Z, Li S, et al. A clustering-based strategy to identify coincidental correctness in fault localization[J]. International Journal of Software Engineering and Knowledge Engineering, 2013, 23 (05): 721-741.
- [125] Unterkalmsteiner M, Gorschek T, Cheng C K, et al. Evaluation and measurement of software process improvement—a systematic literature review[J]. Software Engineering, IEEE Transactions on, 2012, 38 (2): 398-424.
- [126] Moløkken K, Jørgensen M. A review of software surveys on software effort estimation [C]. Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on. IEEE, 2003: 223-230.
- [127] Moin A H, Khansari M. Bug localization using revision log analysis and open bug repository text categorization[M]. Open Source Software: New Horizons. Springer Berlin Heidelberg, 2010: 188-199.
- [128] Moreno-Rivera J M, Navarro E. Evaluation of SPL approaches for WebGIS development: SIGTel, a case study[C]. System Sciences (HICSS), 2011 44th Hawaii International Conference on. IEEE, 2011: 1-10.
- [129] Murgia A, Tonelli R, Marchesi M, et al. Refactoring and its relationship with fan-in and fan-out: An empirical study [C]. Software Maintenance and Reengineering (CSMR), 2012 16th European Conference on. IEEE, 2012: 63-72.
- [130] Murphy Hill E, Parnin C, Black A P. How we refactor, and how we know it[J]. Software Engineering, IEEE Transactions on, 2012, 38(1): 5-18.
- [131] Nakagawa E Y, De Sousa E P M, de Brito Murata K, et al. Software architecture relevance in open source software evolution: a case study[C]. Computer Software and Applications, 2008. COMPSAC'08. 32nd Annual IEEE International. IEEE, 2008: 1234-1239.
- [132] NASA Cost Estimation handbook [OL]. http://www. ceh. nasa. gov/webhelpfiles/Cost_Estimating_Handbook_NASA_2004. htm. 2004.
- [133] Engelhart J, Langbroek P. Function Point Analysis (FPA) for Software Enhancement [M]. Jeddah: NESMA, 2001.
- [134] Niazi M, Wilson D, Zowghi D. A model for the implementation of software process improvement: A pilot study [C]. Quality Software, 2003. Proceedings. Third International Conference on. IEEE, 2003: 196-203.
- [135] Laitenberger O, El Emam K, Harbich T G. An internally replicated quasi-experimental comparison of checklist and perspective based reading of code documents[J]. Software Engineering, IEEE Transactions on, 2001, 27(5): 387-421.
- [136] De Oliveira L B R, Felizardo K R, Feitosa D, et al. Reference models and reference architectures based on service-oriented architecture: a systematic review[M]. Heidelberg: Springer, 2010: 360-367.
- [137] Palomba F, Bavota G, Di Penta M, et al. Detecting bad smells in source code using change history

- information [C]. Automated Software Engineering (ASE), 2013 IEEE/ACM 28th International Conference on. IEEE, 2013: 268-278.
- [138] Park R E. Software size measurement: A framework for counting source statements [R]. CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST, 1992.
- [139] Park Y, Park H, Choi H, et al. A study on the application of six sigma tools to PSP/TSP for process improvement [C]. Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on. IEEE, 2006: 174-179.
- [140] Fabrizio Pastore, Leonardo Mariani ZoomIn. Discovering Failures by Detecting Wrong Assertions [C]. Proceedings of International Conference on Software Engineering. 2015: 66-76.
- [141] Perry D E, Wolf A L. Foundations for the study of software architecture [J]. ACM SIGSOFT Software Engineering Notes, 1992, 17(4): 40-52.
- [142] Peters F, Menzies T, Gong L, et al. Balancing privacy and utility in cross- company defect prediction [J]. Software Engineering, IEEE Transactions on, 2013, 39(8): 1054-1068.
- [143] Pfleeger S L, Wu F, Lewis R. Software cost estimation and sizing methods: issues, and guidelines [M]. California:Rand Corporation, 2005.
- [144] Pinzger M, Gall H C. Dynamic analysis of communication and collaboration in OSS projects [M]. Heidelberg:Springer, 2010: 265-284.
- [145] Qureshi N, Usman M, Ikram N. Evidence in software architecture, a systematic literature review [C]. Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering. ACM, 2013: 97-106.
- [146] Rachatasumrit N, Kim M. An empirical investigation into the impact of refactoring on regression testing [C]. Software Maintenance (ICSM), 2012 28th IEEE International Conference on. IEEE, 2012: 357-366.
- [147] Radjenović D, Herić M, Torkar R, et al. Software fault prediction metrics: A systematic literature review [J]. Information and Software Technology, 2013, 55(8): 1397-1418.
- [148] Rahman F, Khatri S, Barr E T, et al. Comparing static bug finders and statistical prediction [C]. Proceedings of the 36th International Conference on Software Engineering. ACM, 2014: 424-434.
- [149] Ratzinger J, Sigmund T, Vorburger P, et al. Mining software evolution to predict refactoring [C]. Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on. IEEE, 2007: 354-363.
- [150] Rong G, Boehm B, Kuhrmann M, et al. Towards context-specific software process selection, tailoring, and composition [C]. Proceedings of the 2014 International Conference on Software and System Process. ACM, 2014: 183-184.
- [151] Rost D, Naab M, Lima C, et al. Software architecture documentation for developers: A survey [M]. Heidelberg:Springer, 2013: 72-88.
- [152] Saha S K, Selvi M, Buyukcan G, et al. A systematic review on creativity techniques for requirements engineering [C]. Informatics, Electronics & Vision (ICIEV), 2012 International Conference on. IEEE, 2012: 34-39.
- [153] Kollanus S, Koskinen J. Survey of software inspection research [J]. The Open Software Engineering Journal, 2009, 3(1): 15-34.

- [154] Scaachi W. Free/open source software development: Recent research results and methods [J]. *Advances in Computers*, 2007, 69: 243-295.
- [155] Schultis K B, Elsner C, Lohmann D. Architecture challenges for internal software ecosystems: a large-scale industry case study [C]. *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 2014: 542-552.
- [156] Schultis K B, Elsner C, Lohmann D. Architecture challenges for internal software ecosystems: a large-scale industry case study [C]. *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*. ACM, 2014: 542-552.
- [157] Shahin M, Liang P, Babar M A. A systematic review of software architecture visualization techniques [J]. *Journal of Systems and Software*, 2014, 94: 161-185.
- [158] Shahin M, Liang P, Li Z. Do architectural design decisions improve the understanding of software architecture? two controlled experiments [C]. *Proceedings of the 22nd International Conference on Program Comprehension*. ACM, 2014: 3-13.
- [159] Shaw M, Clements P. The golden age of software architecture [J]. *Software*, IEEE, 2006, 23 (2): 31-39.
- [160] Sharafat A R, Tahvildari L. Change prediction in object- oriented software systems: A probabilistic approach [J]. *Journal of Software*, 2008, 3(5): 26-39.
- [161] Shi Q, Chen Z, Fang C, et al. Measuring the Diversity of a Test Set With Distance Entropy [J/OL]. *IEEE TRANSACTIONS ON RELIABILITY*, 2015. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7116633&tag=1>.
- [162] Sivashankar M, Kalpana A M. A framework approach using CMMI for SPI to Indian SME'S [C]. *Innovative Computing Technologies (ICICT)*, 2010 International Conference on. IEEE, 2010: 1-5.
- [163] De Silva L, Balasubramaniam D. Controlling software architecture erosion: A survey [J]. *Journal of Systems and Software*, 2012, 85(1): 132-151.
- [164] SIR[OL]. <http://sir.unl.edu/php/previewfiles.php>.
- [165] Sjøberg D I K, Hannay J E, Hansen O, et al. A survey of controlled experiments in software engineering [J]. *Software Engineering, IEEE Transactions on*, 2005, 31(9): 733-753.
- [166] Soh Z, Khomh F, Guéhéneuc Y G, et al. Towards understanding how developers spend their effort during maintenance activities [C]. *Reverse Engineering (WCRE)*, 2013 20th Working Conference on. IEEE, 2013: 152-161.
- [167] Qinbao Song, Martin J. Shepperd, Carolyn Mair: Using Grey Relational Analysis to Predict Software Effort with Small Data Sets [J]. *IEEE METRICS* 2005: 35.
- [168] Qinbao Song, Martin J. Shepperd, Michelle Cartwright, Carolyn Mair: Software Defect Association Mining and Defect Correction Effort Prediction [J]. *IEEE Transactions on Software Engineering*, 2006, 32 (2): 69-82.
- [169] Stacy K L, Nicholas A K, Letha H E. Source Code Retrieval for Bug Localization using Latent Dirichlet Allocation [C]. *Proceedings of 15th Working Conference on Reverse Engineering*, Antwerp, Belgium, 2008: 155-164.
- [170] Mark Staples, Mahmood Niazi. Two case studies on small enterprise motivation and readiness for CMMI [C]. *Proceedings International Conference on Product Focused Software*, pp. 63-66, ACM (2010).
- [171] Stefan Biffl, et al. Investigating the Defect Detection Effectiveness and Cost Benefit of Nominal Inspection

- Teams[J]. IEEE Transactions on Software Engineering . 2003.
- [172] Stefan Biffl, et al. A family of experiments to investigate the effects of groupware for software inspection [J]. Automated Software Engineering. 2006.
- [173] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, David Redmiles. Social Barriers Faced by Newcomers Placing Their First Contribution in Open Source Software Projects[C]. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, 2015: 1379-1392.
- [174] Stolee K T, S Elbaum. Exploring the use of crowdsourcing to support empirical studies in software engineering [C]. Proceedings International Symposium on Empirical Software Engineering and Measurement. 2010.
- [175] Ting Su, Zhouhai Fu, Geguang Pu, Jifeng He, Zhendong Su. Combining Symbolic Execution and Model Checking for Data Flow Testing[C]. Proceedings of International Conference on Software Engineering. 2015 : 654-665.
- [176] Muhammad Sulayman, Emilia Mendes. An extended systematic review of software process improvement in small and medium Web companies[C]. Proceeding Internation Conference of Evaluation & Assessment in? Software? Engineering, 2011 : 134-143.
- [177] R B Svensson, M Host, B Regnell. Managing quality requirements: A systematic review[C]. Proceedings of 36th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), 2006: 261-268.
- [178] Hee Beng Kuan Tan, Yuan Zhao, Hongyu Zhang. Estimating LOC for information systems from their conceptual data models[C]. Proceedings of International Conference on Software Engineering. 2006: 321-330.
- [179] Hee Beng Kuan Tan, Yuan Zhao, Hongyu Zhang. Conceptual data model-based software size estimation for information systems [J]. ACM Transactions on Software Engineering Methodology. 2009 : 19 (2) : 1-37.
- [180] Taneja K, D Dig, T Xie. Automated detection of API refactorings in libraries[C]. in Proceedings of IEEE/ACM international conference on Automated software engineering. 2007.
- [181] A Tang, M A Babar, L Gorton, J Han. A survey of architecture design rationale[J]. Journal of systems and software, 2006 : 79(12) : 1792-1804.
- [182] Valerio Terragni, Shing- Chi Cheung, Charles Zhang. RECONTEST: Effective Regression Testing of Concurrent Programs [C]. Proceedings of International Conference on Software Engineering. 2015 : 246-256.
- [183] D Tofan, M Galster, P Avgeriou. Difficulty of architectural decisions-a survey with professional architects [C]. Proceedings of the 7th European Conference on Software Architecture (ECSA), Montpellier, France, 2013 : 192-199.
- [184] D Tofan, M Galster, P Avgeriou, W Schuitema. Past and future of software architectural decisions - A systematic mapping study[J]. Information and Software Technology, 2014 : 56(8) : 850-872.
- [185] N Tsantalis, A Chatzigeorgiou, G Stephanides. Predicting the probability of change in object- oriented systems [J]. IEEE Transactions on Software Engineering, 2005 , 31(7) : 601-614.
- [186] Jingxuan Tu, Lin Chen, Yuming Zhou, Jianjun Zhao, Baowen Xu. Leveraging Method Call Anomalies to Improve the Effectiveness of Spectrum-Based Fault Localization Techniques for Object-Oriented Programs [C]. QSIC 2012 : 1-8.

- [187] UKSMA. MK II Function Point Analysis: Counting Practices Manual, Version 1.3.1 [C/OL]. United Kingdom Software Metrics Association (UKSMA), 1998. <http://www.uksma.co.uk/public/mkIIR131.pdf>.
- [188] F Vogelezang. COSMIC full function points the next generation of functional sizing [C/OL]. In: Software Measurement European Forum SMEF 2005, 2005.
- [189] L Votta, et al. Does every inspection need a meeting? [J/OL]. ACM Software Engineering Notes, 1993.
- [190] F Walkerden, R Jeffery. An Empirical Study of Analogy-based Software Effort Estimation [J]. Journal of Empirical Software Engineering, 1999; 4(2): 135-158.
- [191] Yong Wang, Qinbao Song, Stephen G. MacDonell, Martin J. Shepperd, Junyi Shen. Integrate the GM(1, 1) and Verhulst Models to Predict Software Stage Effort [J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009; Part C. 39(6): 647-658.
- [192] Wang S, Zhang W, Yang Y, Wang Q DevNet. Exploring Developer Collaboration in Heterogeneous Network of Bug Repositories [C]. Proceedings ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM'13), Oct 2013, Maryland, USA. 2013; 193-202.
- [193] J Wen, S Li, Z Lin. Systematic Literature Review of Machine Learning Based Software Development Effort Estimation Models [J]. Information and Software Technology, 54 (1): 41-59, 2012.
- [194] Michael W Whalen, Suzette Person, Neha Rungta, Matt Staats, Daniela Grijincu. A Flexible and Non-intrusive Approach for Computing Complex Structural Coverage Metrics [C]. Proceedings of International Conference on Software Engineering, 2015; 506-516.
- [195] I. Wieczorek, M Ruhe. How Valuable is company-specific Data Compared to multi-company Data for Software Cost Estimation? [C]. Proceedings of the 8th IEEE International Symposium on Software Metrics, 2002; 237-246.
- [196] B J Williams, J C Carver. Characterizing software architecture changes: A systematic review [J]. Information and Software Technology, 2010; 52(1): 31-51.
- [197] Chu-Pan Wong, Yingfei Xiong, Hongyu Zhang, Dan Hao, Lu Zhang, Hong Mei. Boosting Bug-Report-Oriented Fault Localization with Segmentation and Stack-Trace Analysis [C]. Proceedings International Conference on Software Maintenance and Evolution. 2014; 181-190.
- [198] Edmund Wong, Lei Zhang, Song Wang, Taiyue Liu, Lin Tan. DASE: Document-Assisted Symbolic Execution for Improving Automated Software Testing [C]. Proceedings of International Conference on Software Engineering, 2015; 620-631.
- [199] 伍书剑. 基于系统动力学技术的缺陷预测模型及经验研究 [D]. 中国科学院研究生院, 2010.
- [200] Xie X, Zhang W, Yang Y, Wang Q. DRETOM: developer recommendation based on topic models for bug resolution [C]. Proceedings of the 8th International Conference on Predictive Models in Software Engineering (PROMISE'12). September 21-22, 2012, Lund, Sweden. ACM, 19-28.
- [201] Xu J, G Madey. Exploration of the open source software community [C]. Proceedings of North American Association for Computational Social and Organizational Science (NAACOS). 2004. Pittsburgh, PA, USA.
- [202] Xuan J, Jiang H, Ren Z, Zou W. Developer Prioritization in Bug Repositories [C]. Proceedings of 34th International Conference on Software Engineering. June. 2012, 25-35.
- [203] Yang J, L A Adamic, M S Ackerman. Crowdsourcing and knowledge sharing: strategic user behavior on tasken [C]. Proceedings of the 9th ACM conference on Electronic commerce (EC'08). New York,

NY, USA.

- [204] Yu L, S, Ramaswamy. Mining cvs repositories to understand open-source project developer roles [C]. Proceedings of the 4th International Workshop on Mining Software Repositories. 2007: 8.
- [205] Yuan Z, Yu L, Liu C. Bug prediction method for fine-grained source code changes [J/OL]. Ruan Jian Xue Bao/Journal of Software, 2014, 25(11) : 2499-2517 (in chinese). <http://www.jos.org.cn/1000-9825/4559.htm>. [doi: 10.13328/j.cnki.jos.004559].
- [206] Yuen M C, I King, K S Leung. Task Matching in Crowdsourcing [C]. Proceedings of the 4th International Conference on Cyber, Physical and Social Computing. 2011: 409-412.
- [207] Zhang W, Yang Y, Wang Q. Network analysis of OSS evolution: an empirical study on ArgoUML project [C]. In Proceedings of the 12th International Workshop on Principles of Software Evolution and the 7th annual ERCIM Workshop on Software Evolution, Szeged, Hungary, 2011: 71-80.
- [208] Zhang, He, Muhammad Ali Babar, Paolo Tell. Identifying relevant studies in software engineering [J]. Information and Software Technology 53. 6 (2011) : 625-637.
- [209] Hongyu Zhang, Liang Gong, Steven Versteeg: Predicting bug-fixing time: an empirical study of commercial software projects [C]. Proceedings of International Conference on Software Engineering. 2013: 1042-1051.
- [210] Yuming Zhou, Hareton Leung. Empirical Analysis of Object-Oriented Design Metrics for Predicting High and Low Severity Faults [J]. IEEE Transactions on Software Engineering. 2006: 32(10) : 771-789.
- [211] Zhou Y, Leung H. Predicting object-oriented software maintainability using multivariate adaptive regression splines [J]. Journal of Systems and Software, 2007: 80(8) : 1349-1361.
- [212] Zhou Y, Leung H K, N Xu B. Examining the Potentially Confounding Effect of Class Size on the Associations between Object-Oriented Metrics and Change-Proneness [J]. IEEE Transactions on Software Engineering. 2009: 35(5) : 607-623.
- [213] Zhou J, Zhang H, David L. Where should the bugs be fixed? More accurate information retrieval-based bug localization based on bug reports [C]. Proceedings of 34th International Conference on Software Engineering, Zurich, Switzerland, 2012: 14-24.
- [214] Zhu X, Song Q, Sun Z. Automated Identification of Change-Prone Classes in Open Source Software Projects [J]. JSW, 2013: 8(2) : 361-366.
- [215] Zimmermann T, et al. Mining Version Histories to Guide Software Changes [C]. Proceedings of 26th International Conference on Software Engineering. 2004: 429-445.
- [216] Zimmermann T, A Zller, P Weigerber, et al. Minin version histories to guide software changes [J]. IEEE Transactions on software Engineering, 2005, 31(6) : 429-445.
- [217] Zimmermann T, Nagappan N. Predicting Defects using Network Analysis on Dependency Graphs [C]. Proceedings of the 30th International Conference on Software Engineering, 2008: 531-540.
- [218] Zimmermann T, Nagappan N, Gall H. Cross-project defect prediction: a large scale experiment on data vs. domain vs. process [C]. Proceedings of the 7th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering, 2009: 91-100.
- [219] Kitchenham B, Dybå T, Jørgensen M. Evidence-based Software Engineering [C/OL]. Proceedings of 26th International Conference on Software Engineering (ICSE), 2004: pp. 273-281.
- [220] Kitchenham B. Guidelines for Performing Systematic Literature Reviews in Software Engineering [R]. Software Engineering Group, School of Computer Science and Mathematics, Keele University, and

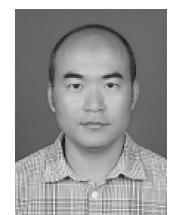
- Department of Computer Science, University of Durham (2007).
- [221] Kitchenham B. Procedures for Undertaking Systematic Reviews. Computer Science Department [R], Keele University and National ICT Australia (2004).
- [222] Hossain E, Babar M. A, Paik H. Using Scrum in Global Software Development: A Systematic Literatur Review [C]. Proceedings of the 4th International Conference on Global Software Engineering. IEEE Press, Los Alamitos, pp. 195-204.
- [223] Beijun Shen, Ruan Tong. A Case Study of Software Process Improvement in a Chinese Small Company [C]. Proceedings of International Conference on Computer Science and Software Engineering (CSSE 2008). Dec. 2008, Wuhan, China, PP. 609-612.
- [224] E Guzman, W Maalej. How do users like this feature? A finegrained sentiment analysis of App review [C]. Proc. 22nd IEEE International Requirements Engineering Conference (RE2014), Karlskrona, Sweden, August 2014, pp. 153-162.
- [225] N Chen, J Lin, S Hoi, X Xiao, B Zhang. AR-miner: mininginformative reviews for developers from mobile App marketplace [C]. Proceedings of 36th International Conference on Software Engineering (ICSE2014), Hyderabad, India, Jun. 2014, pp. 767-778.
- [226] Ke Li, Junchao Xiao, Yongji Wang, Qing Wang. Analysis of the Key Factors for Software Quality in Crowdsourcing Development: An Empirical Study on TopCoder. com [C]. Proceedings of the 37th Annual International Computer Software & Applications Conference(Compsac2013), July 22-26, 2013.

作者简介

王青 女，博士，研究员、博士生导师，现任中国科学软件研究所副总工程师，主要研究方向包括：软件过程、软件质量保障、需求工程、知识工程等。目前主要的社会兼职为中国电子学会云计算专委会委员，全国信息技术标准化委员会软件质量测试工作组（SAC/TC28/SC7/WG1）副组长，以及CMMI认证授权的主任评估师。她主持和承担了多项国家重点/重大项目和国内外重大合作项目，获得国家及省部委科技进步奖10余次。还曾应邀担任多个国际会议的程序委员会主席或委员，同时担任ESEIW/ESEM 2015 (ACM 国际经验软件工程与度量会议) 大会主席，也是 IST、JSS 等国际期刊审稿人，近年有 5 本论/编著、100 余篇论文在国际、国内重要学术刊物以及国际会议上发表。也曾在中国科学院大学主讲《高级软件工程》课程。详细履历见 <http://itechs.icscas.ac.cn/cn/education/wq.htm>。

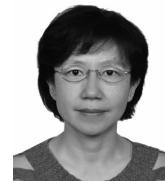


张贺 哲学博士 (Ph. D)，南京大学软件学院教授，博士生导师，国际软件工程研究联盟 (ISERN) 成员、南京大学代表，中国计算机学会高级会员、软件工程专委会委员。毕业于澳大利亚新南威尔士大学 (UNSW)，以最优成绩获博士学位 (Ph. D)。师从世界顶级软件工程专家 (澳) Ross Jeffery 教授和 (英) Barbara Kitchenham 教授。在欧洲和澳洲从事软件工程研究与实践 10 余年。先后在爱尔兰国家软件工程研究中心



(Lero) 任研究员，在澳大利亚国家信息与通信科学院（NICTA）任资深研究员，并在（爱）利默里克大学（UL）信息技术系和（澳）新南威尔士大学（UNSW）计算机科学与工程系任特聘讲师。也曾负责主持多个爱尔兰、澳大利亚、中国等国家级科研基金项目。近年来，著有英文专著一部，并在国际软件工程大会（ICSE）和 Empirical Software Engineering、Information and Software Technology、Journal of Systems and Software、IEEE Transactions on Service Computing、Journal of Software: Evolution and Process 等国际主要软件工程期刊和会议上发表论文 90 余篇（约 40% 为第一作者），其中 8 篇获最佳论文奖。2012 年入选“登峰计划”（A-），2013 年起，任教于南京大学。

张 莉 北京航空航天大学计算机学院/软件学院教授，博士生导师，中国计算机学会软件工程专委会委员及教育专委副主任委员，教育部软件工程专业教学指导委员会委员，全国工程专业学位研究生教育指导委员会软件工程领域协作组组长。她的研究兴趣是：软件工程（系统设计、软件重用、领域/系统建模、经验软件工程等）。



梁 鹏 武汉大学软件工程国家重点实验室教授、博士生导师，主要研究方向为软件体系结构、知识驱动的软件工程。



周明辉 北京大学信息科学技术学院副教授，主要研究兴趣是：通过挖掘软件开发支持工具（version control system、issue tracking system、mailing list 等）记录的历史数据，研究软件产品和程序员及其工作文化之间的关系（尤其是开源项目）。在软件工程领域顶级国际期刊及 TSE、ICSE 和 FSE 会议等发表 40 多篇论文，获 FSE2010 的 ACM SIGSOFT 杰出论文奖和 COMPSAC 2012 最佳论文奖。还多次担任国际会议 PC，如 FSE 2014 Research Demo PC co-chair 和 Internetware 2014 PC co-chair 等。



知识型服务计算

CCF 服务计算专委会

摘要

随着云计算、大数据、社交网络、移动互联网、物联网等新兴技术的出现，在面向服务的复杂生态系统中，越来越强调以更加智能化、个性化和自动化的处理方式为用户提供智慧服务，这使得知识型服务计算的研究变得越来越重要。本文分析了知识型服务计算的国内外研究进展，并对其发展前景进行了展望。

关键词：知识型服务计算，服务发现，服务推荐，服务组合，业务流程管理

Abstract

With the rapid development of cloud computing, big data, social network, mobile internet, and Internet of Things, it is emphasized to provide services to users in a more intelligent, personalized and automatic way in service-oriented complex ecosystems, which makes knowledge-based services computing increasingly important. This report surveys the recent progress of knowledge-based services computing, makes a comparison on these efforts, and outlines some development trends in this research domain.

Keywords: knowledge-based services computing, service discovery, service recommendation, service composition, business process management

1 引言

在互联网逐渐高度发达的今天，实现“互联网+”的核心技术之一——服务计算也正经历着变革性发展。作为弥合业务服务和IT服务间鸿沟的桥梁，服务计算的核心思想是，将互联网中分布的服务作为向用户提供所需功能的基本单元，在对用户需求与服务资源的能力加以匹配的基础上，进行服务的重用、组合、定制以及测试，从而快速构建能随需应变、松散耦合的网络服务应用。

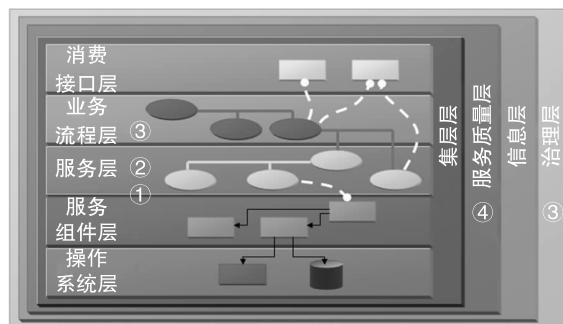
近年来，云计算、大数据、社交网络、移动互联网、物联网等新兴技术的出现，进一步推动了服务计算的发展。随着多源异构服务的规模化剧增、服务资源的持续演化，以及互联网海量用户个性化需求的不断涌现，在面向服务的复杂生态系统中越来越强调以更加智能化、个性化和自动化的处理方式为用户提供服务。为此，如何挖掘和利用业务领域与计算资源的大数据中所蕴含的知识支持上述服务就显得更加重要。有两类以人为主体的知识：一是面向业务领域进行服务资源规模化定制、组织、管理的领域专家知识；二是面向领域及主题的社会化涉众群体对服务系统效能的认识，并往往以显性的“问

答卷”“评论”“标签”“事件报告”“服务日志”等流式“密集型数据”所隐含的价值服务知识。这两类知识通过定性、定量、感性计算的语义综合集成迭代式整合形成“望你所为”的知识，来引导、影响、预测并约束与激励实际的服务计算系统。这两类知识及其与实际的服务计算系统之间的互动与演化生长，正是知识型服务计算工作的关注重点。

从科学发展的规律来看，无论是自然科学的计算机科学与技术、自动化、认知科学、复杂性科学等学科，还是社会科学中的系统工程与管理科学、科学哲学等学科的发展，今天几乎同时出现了一种聚焦“知识型工作”的“异途同归”趋势。其本质原因在于互联网的发展，已能使人文（以人为本）的知识几乎可以瞬间传播并获取、瞬间影响可以遍及整个网络信息空间、人人时时都有参与和发挥作用的机会。因此，以面向数据密集型科学发现的“第四范式”（The Fourth Paradigm^[1]）为代表的 knowledge型 work 将发挥越来越重要的作用，即在面向领域主题的涉众群体知识逐步丰富（认知过程）条件下，再过渡到传统的第二范式（模型算法归纳）、第三范式（模拟仿真推演）的工作模式，将成为知识型工作的科学研究元范式。从传统的服务计算到知识型服务计算的创新发展，是其中的一种范例。

在这一背景和发展趋势下，知识型服务计算的研究变得越来越重要。将知识型工作在服务计算中进行广泛和深入的融合应用，充分利用机器学习、本体和概念建模、业务规则推理等各种知识工程相关的方法和技术，在面向服务的生态系统中对大数据进行知识挖掘和应用，成了当今服务计算的重要需求。

国际开放群组（The Open Group）提出的 SOA RA 参考架构^[2]是在服务计算领域中得到广泛认可的参考架构。如图 1a 所示，SOA RA 定义了在设计 SOA 解决方案或定义基于 SOA 的企业结构时需要考虑的 9 个重要层次，包括 5 个水平层：操作系统层、服务组件层、服务层、业务流程层和消费接口层，以及 4 个垂直层：治理层、信息层、服务质量层和集成层。水平层主要与 SOA 解决方案的功能特征直接相关，而垂直层则体现为支持跨功能层的横切关注点。在本报告中，我们对知识型服务计算的分析主要是从基于知识的服务发现与推荐、基于知识的服务组合、基于知识的业务流程管理，以及基于知识的服务质量管理和预测等几个方面展开（如图 1b 所示），图 1a 中也描述了这几部分内容在 SOA RA 中所处的位置，其中，基于知识的服务发现与推荐主要与服务层和服务组件层相关；基于知识的服务组合主要与服务层相关；基于知识的业务流程管理主要与业务流程层和治理层相关；而基于知识的服务质量管理与预测则主要与服务质量层相关。



a) SOA RA 参考架构层次图（引自文献[2]）

③基于知识的业务流程管理	④基于知识的服务质量管理与预测
②基于知识的服务组合	
①基于知识的服务发现与推荐	

b) 知识型服务计算的核心内容

图 1 知识型服务计算的核心内容及其与 SOA RA 的关系

本报告从上述 4 个方面入手，通过对国内外相关研究的介绍和分析，展示当前知识型服务计算的发展现状，进而对知识型服务计算的未来发展趋势进行展望。

2 国际研究现状

2.1 基于知识的服务发现与推荐

(1) 服务发现

如图 2 所示，服务发现方法可根据服务注册信息的部署方式不同分为集中式 (Centralized)^[3] 和分布式 (Decentralized)^[4-21]，其中分布式方法又可细分为结构化的 P2P、非结构化的 P2P 和混合式三类。纵轴主要按支持非描述逻辑、描述逻辑和混合型分为三类。

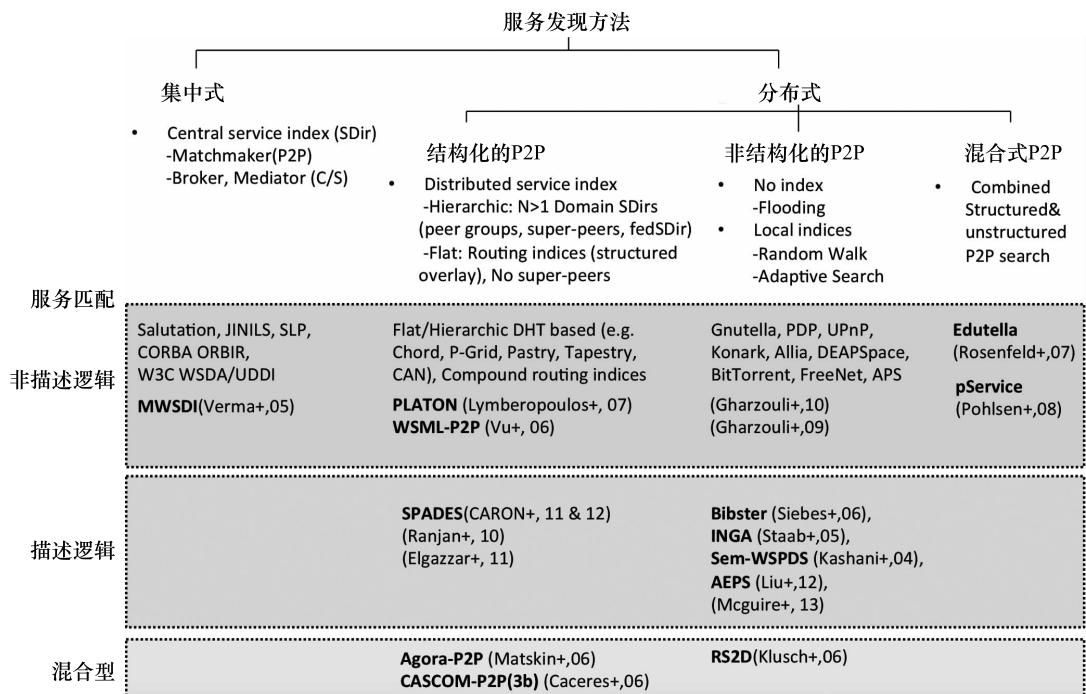


图 2 国际现有主要服务发现方法

针对支持非描述逻辑的方法，国际上对于服务发现方法在集中式和分布式两类中的研究都较为成熟。文献 [11] 为 Web 服务自动组合提出了一个分布式架构，该架构实现了一个基于传染性发现算法的分布式解决方案，在无结构的 P2P 网络中进行语义的 Web 服务发现。文献 [3] 提出了一个在多个注册中心中可扩展、高性能的 Web 服务发布和发现环境，主要利用基于本体论的方法实现基于域对所有 Web 服务进行分类。在支持描

述逻辑和混合型的两类研究中，基于集中式的服务发现方法研究工作开展得较少，主要集中于分布式方法。文献 [19] 提出一个自适应高效的 P2P 搜索方法，该方法用于基于大量社交行为模式的 SOA 上的可靠服务整合。

(2) 服务推荐

目前主要的服务推荐方法可由如图 3 所示的两个维度展现。在水平方向上，按照推荐方法的类别，分别将服务推荐方法分为三类：基于内容 (content-based)^[22-27]、基于协同过滤的邻近关系 (neighbor-based)^[28-30]，以及情境感知 (context-aware) 的服务推荐^[31-38]。上述三种推荐类型又分别细分为两类。在垂直方向上，根据推荐过程使用的信息分为三类：使用服务的功能性信息 (functional)、使用服务的非功能性信息 (non-functional)，以及同时使用二者信息 (hybrid)。

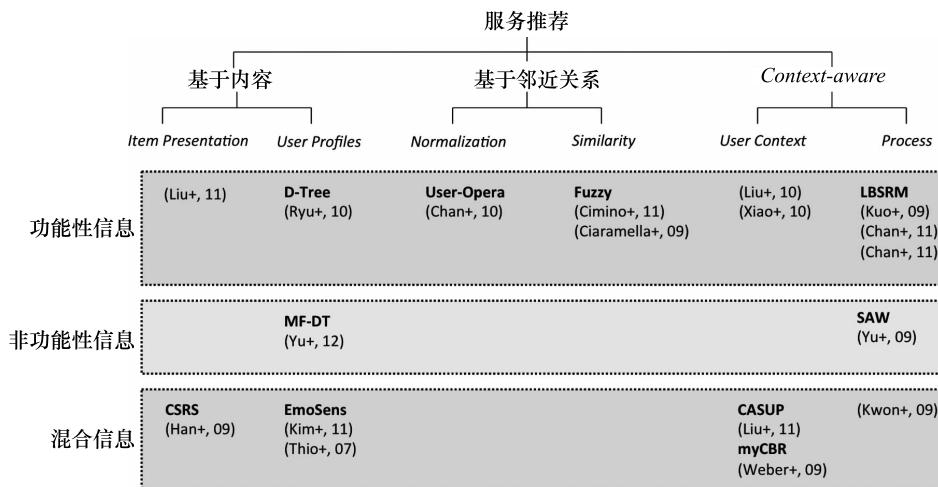


图 3 国际现有主要服务推荐方法

从垂直方向上看，国际上现有的服务推荐方法基于服务功能性信息在主流的三类推荐方法中均取得了一定的研究进展。文献 [34-35] 提出通过请求服务周围的流程片段获取请求服务的指定组合情境，从而推荐在已存在的服务组合中组合情境与给定的片段情境最匹配的服务。文献 [22] 提出的模型利用决策树对用户进行聚类，并通过遍历决策树分析新来的用户以减少进行基于内容推荐过程中的分析时间，提高推荐效率。在使用非功能性信息方面，研究成果较为不足，尤其是在基于协同过滤的邻近关系的服务推荐领域。在基于内容的服务推荐方法领域，文献 [24] 利用矩阵分解 (MF) 结合决策树学习解决了协同过滤中用户的冷启动问题。在同时使用服务的功能性信息和非功能性信息方面，针对基于内容和情境感知的推荐方法的研究较为热门。文献 [27] 提出的推荐框架利用 Web 服务配置文件和客户信息估计客户端性能，以便推荐最合适的服务提供商。

2.2 基于知识的服务组合

服务组合是服务计算领域的经典问题，针对特定客户（群）的明确/隐式需求，从

可用候选服务集中选择一组服务并根据它们之间的语义/接口/数据依赖关系连接起来，形成一个组合服务以满足客户需求。服务组合问题得到研究者的广泛关注，形成了相对成熟的理论与方法体系，请参阅相关的综述论文^[39-41]。

知识在服务组合问题的求解中发挥着重要作用。若将每个候选服务看作一个“知识源”、将客户初始状态看作“初始知识”、将客户期望状态看作“目标知识”、将所处领域和环境的约束看作“领域知识”，那么服务组合问题就可看作是“知识源选择问题”或称“组件集识别问题”，寻求从初始知识到目标知识的映射路径，将所选服务的行为加以综合，得到满足约束的优化组合方案。服务组合本质上是分布式的知识和行为的组合，可称为基于知识的服务组合。

在宏观层面，服务组合所涉及的知识以及建立在各类知识上的研究问题分为以下几个类别。

1) 从候选服务的角度：各服务自身蕴含着丰富的信息，这是最低层次的知识，用于对候选服务的性质和能力进行刻画，将它们作为“知识源”对外呈现，涉及对服务的行为、接口、质量（QoS）、合约（Contract）、隐私，以及不同服务间的相关性/依赖性/相似性/竞争性等方面表示，是服务组合问题的基本输入。

2) 从需求的角度：对客户的显式和隐式需求进行知识化表示，涉及客户个性化特征、群体客户的共性特征、客户初始状态与期望状态、客户对状态转换过程的约束。考虑到服务的个性化特征，对客户个性化需求特征和模式的挖掘是当前研究的热点之一。考虑到服务提供者的成本有效性，分析多用户需求之间的相似性以大规模个性化的方式提供组合方案，也成为当前研究的热点。

3) 从环境的角度：不管是各候选服务还是客户，其所处的环境均是动态变化的；此类知识用于对环境动态特性进行刻画，称之为上下文相关（Context-aware）的知识。环境的变化使服务组合不再仅被看作设计阶段的任务，还要考虑执行阶段的客户需求变化、客户上下文变化、候选服务的上下文变化等因素。

4) 从时间的角度：服务组合并非针对孤立的单次需求，在当前时刻之前所进行的服务组合历史数据中隐含着丰富的知识，可作为后续组合的先验性知识（如典型的需求模式、服务组合模式、组合案例），可帮助对组合过程和结果的性能做出有效预测。

5) 从知识可见性的角度：以上各类知识既有显式的、可公开发现和分享的知识，有局部可见可用的知识，也有隐式的、需要进行探索和预测的知识；有相对准确的知识，也有相对模糊或不准确的知识。如何探索和预测隐式的知识、如何将模糊的知识进行精细化，以保障知识对组合效能的帮助，也是当前的关注点之一。

基于知识的服务组合方法可分为以下几个类别。

1) 基于组合优化的服务组合方法：将服务组合问题看作多维度、多目标、多重选择的背包问题（MMMKP），利用运筹学中的整数规划等经典方法发现全局最优解。为降低时间复杂度，基于各类启发式规则和知识的局部优化算法、局部优化和全局优化相结合的优化算法相继提出，其典型策略是减小搜索空间（如 Skyline 方法）、将全局约束分解为局部约束、在多个粒度层次进行分层组合等。基于仿生学的各类演化算法也被广泛用

于服务组合求解。

2) 基于人工智能规划的服务组合方法^[61-62]: 将各候选服务看作知识源, 其行为对客户状态产生特定的变迁, 利用特定的 AI 规划器寻求从客户初始状态到客户期望的目标状态之间的服务网络, 同时满足服务质量、上下文等方面的约束。包括基于图规划的组合算法、基于分层任务网络和目标分解的规划算法^[65-66]、基于逻辑推理的规划算法、并行化图规划算法等。

3) 基于多智能体协商的服务组合方法^[69-70]: 不再认为服务组合是由某一方独立完成的过程, 而是将各候选服务看作是具有自省能力的智能体, 基于特定的多智能体协调协议, 根据组合任务的目标和自己所提供的服务行为来决定与其他智能体如何协调和决策、如何验证组合结果的合法性, 从而实现分布式的、去中心化的规划过程。

4) 基于历史数据挖掘的服务组合方法^[51,73]: 充分利用客户的历史需求以及相应的历史组合结果, 对其进行分类和聚类, 发现其中蕴含的潜在模式(需求模式、服务组合模式, 以及二者之间的映射关系), 作为先验知识, 既降低了搜索空间, 也提升了满足客户潜在需求的程度。

5) 基于人机交互的半自动化和探索式服务组合方法^[75-76]: 不追求服务组合过程的全自动化, 认为客户需求无法一次性完全表述, 故需在组合过程中逐渐发现潜在需求及其变化, 对客户做出针对性的引导和推荐, 由客户做出组合过程中的选择和决策, 逐步达到目标。

6) 基于社会智能的服务组合方法^[77-78]: 利用社会计算的相关方法, 考虑客户之间形成的社交网络, 用户可与他人分享服务组合的结果和相应的知识, 根据他人分享的知识来构造满足个人需求的组合方案。该类方法将社交网络作为传播知识的渠道, 侧重于对群体智能的应用。

7) 基于大规模定制的服务组合方法^[79-80]: 不再是针对单一客户, 而是针对群体客户, 利用客户的相似性对多需求分组, 利用少量的组合服务来满足多需求, 或者利用具有可定制能力的组合服务网络来派生出满足各需求的组合方案, 从而在服务组合的成本有效性和个性化需求满足程度之间达成折中。

知识的表示和推理是基于知识的服务组合方法的重要基础, 主要来自于人工智能领域。广泛使用的领域知识表示方法有本体^[65,81-82]、PDDL、标签、规则、各类描述逻辑^[70,72]等。

2.3 基于知识的业务流程管理

基于知识的业务流程管理主要包括两个方面的内容: 一是业务流程知识的挖掘, 其内涵是基于数据挖掘技术, 挖掘信息系统遗留下来的各类流程事件日志, 得到多维的业务流程模型; 二是知识在业务流程管理中的应用, 其内涵是利用领域本体、特征模型、调查问卷等各种知识工程相关方法与技术, 对业务流程进行智能化、自动化的管理与应用。知识在业务流程管理中的应用涉及业务流程管理研究的各方面。图 4 是基于知识的

业务流程管理的频谱图。其中，圆矩形描述了近年来国际国内学术界与工业界在基于知识的业务流程管理方面的一些研究工作，经过近十年发展，这些研究一直被持续关注，并取得重要进展。

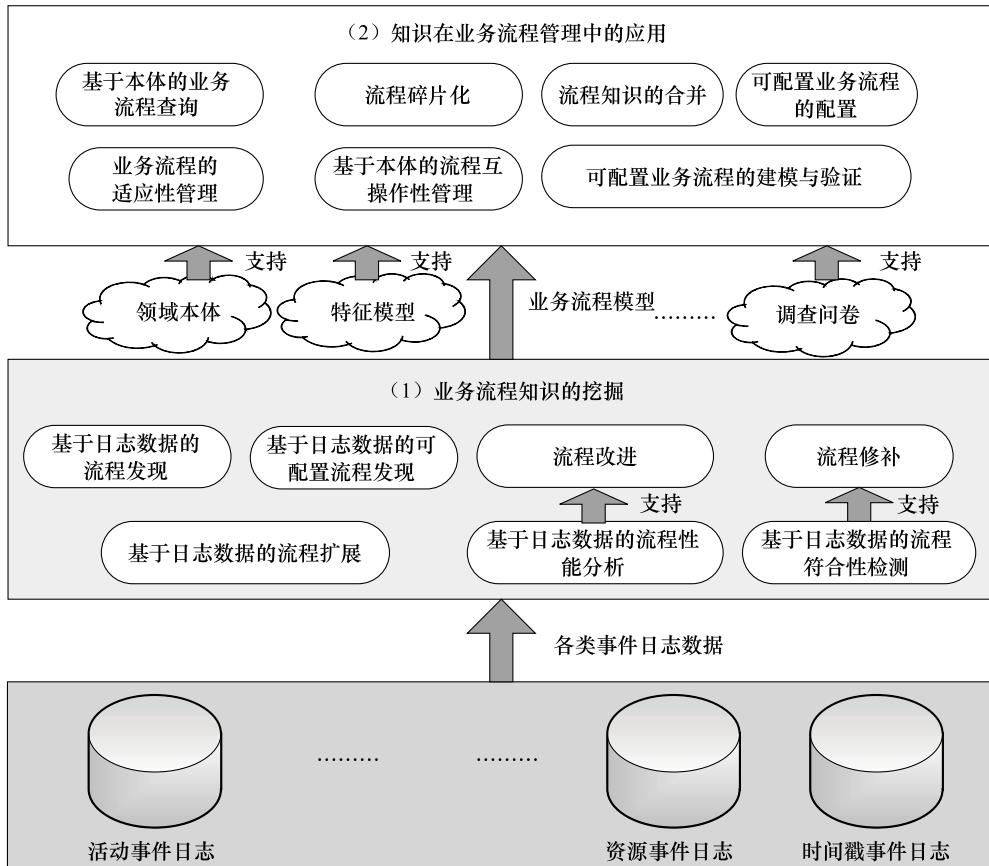


图 4 基于知识的业务流程管理

(1) 业务流程知识的挖掘

现代信息系统与业务流程管理的结合日益紧密。由于大量的事件日志数据可被信息系统记录下来，这为流程挖掘技术的发展奠定了坚实基础。流程挖掘的目标是基于数据挖掘技术，利用这些事件日志数据来抽取、发现业务流程相关知识。一般来讲，流程挖掘方法将利用信息系统执行过程中所记录的有关业务流程活动的序列事件及其对应的流程实例信息来挖掘流程。有些流程挖掘方法还会利用一些附加的日志信息，如，实际执行或者初始化流程活动的人或资源、事件的时间戳，以及与事件一同被记录的数据元素等，来挖掘多维的业务流程模型。

流程发现方法是一类典型的流程挖掘算法，这类算法通常不需要使用任何先验 (a-priori) 信息，而是仅仅利用事件日志来挖掘出一个业务流程模型。业务流程管理领域知名学者 Wil van der Aalst 提出了 α 算法，通过分析事件日志生成一个可以解释记录在日志中的行为的 Petri 网^[84-85]。如果其中的时间日志还包含有关资源的信息， α 算法还可以

用于挖掘流程资源相关的知识，比如利用 α 算法挖掘出某企业里协同工作的人们所形成的社交网络。Buijs 等在文献 [86] 中分析并比较了四种主流的可配置业务流程的挖掘方法，与 Wil 所提出的 α 算法不同的是，Buijs 所分析的挖掘方法可以挖掘出多个相似的流程变体。

流程的知识挖掘技术还可以用来对流程进行符合性检查，通过比较日志数据和流程模型来判定流程的建模行为与流程的可观察行为是否一致。人们可以通过“回放”一个流程模型的历史数据来观察可观测行为与流程模型在何处不一致。Munoz-Gama 等在文献 [87] 中提出了一种针对大尺度流程模型的符合性检测方法。首先，将具有大尺度的流程模型通过流程分解技术分割成若干个单输入 - 单输出的子流程；然后，分别对每个子流程的事件日志同时进行并行分析，从而提高流程符合性分析的效率。对流程符合性检查所产生的诊断结果可以作为流程修补研究工作的输入。流程修补是指通过挖掘事件日志中的知识，使得业务流程模型更加适应业务实际运行的情况。Fahland 等在文献 [88] 中提出了一种比较经典的流程修补算法。首先，利用已有的符合性检测器将给定的流程模型与日志中的流程轨迹进行对齐；然后，基于以上信息，将流程的日志分解成为若干个非拟合的子路径。接下来，对于每个子路径构造出对应的子流程，并在合适的位置上加入原来的流程模型中，从而完成对原给定流程模型的修补。Beest 等在文献 [89] 中基于人工智能 AI 规划方法给出了另一种流程修补的方法。首先，流程的每一个活动都被打上前置条件和后置条件的标签。当流程模型的某一个实例由于资源异常而无法正常运行时，该方法可以定位出资源发生异常的位置，然后通过 AI 规划自动生成一段新的具有一致业务目标的流程片段来代替无法运行的那部分流程片段，从而完成对发生异常的流程实例的修补。

通过利用一些带有时间戳的事件日志，流程挖掘技术还可以用来诊断业务流程模型的性能。通过回放给定流程带有时间戳的事件日志，人们可以测量两个相邻活动间的时间间隔，从而发现流程执行中的瓶颈，以支持流程预测（即运行这个流程实例预计还需要多长时间）和流程推荐（哪一个任务使流程整体成本最小化）^[90]。这些挖掘出来的知识可用来支持流程改进，从而降低流程成本及响应时间，帮助企业改进已有的旧流程^[91]。

通过对流程事件日志进行知识挖掘还可以扩展模型。如上文所述，事件日志除了执行的活动序列以外，还可以是一些标记了附加信息的事件，比如，执行事件的资源、事件的时间戳，以及记录在事件中的数据元素。流程模型扩展就是通过挖掘事件日志中的这些附加知识来不断地扩充、丰富初始只含有控制流的业务流程模型。Rozinat 等在文献 [92] 中提出了从带数据标注的事件日志中挖掘出流程模型的方法，并用着色 Petri 网表示挖掘出来的模型。

（2）知识在业务流程管理中应用

业务流程模型是一种重要的企业遗留知识，需要能够很好地复用。业务流程仓库是用来存储业务流程模型的存储库。为了更好地支持业务流程模型的复用，人们使用本体标注流程模型及其某些属性，使得流程模型可以自动地发布、发现并执行。例如，用 Web 本体语言 OWL 描述一个复杂服务内部的流程模型，形成 OWL-S Process Model^[93]、

Semantic EPC^[94] 等语义标注的业务流程模型。Hepp 等将语义网技术和业务流程管理相结合, 为业务流程模型提供一种合适的知识表示形式, 形成所谓的语义业务流程管理方法, 从而支持基于逻辑表达式的流程模型的查询^[95]。

大规模流程仓库中存储的业务流程模型数量往往较为庞大, 导致可能存在一些重复的流程片段, 造成流程模型存储时的冗余。Dumas 等提出了在大规模流程仓库中快速检测可复用的流程片段的方法^[96]。首先, 将流程仓库中的流程模型通过单输入 - 单输出技术进行分割; 然后, 对分割之后的流程片段建立索引, 并将流程片段转换成标准代码并加以注册, 从而有效地提高了大规模流程仓库的存储效率。Eberle 等提出将流程碎片作为构建流程的基本模块, 并在流程设计和运行时加以复用^[97]。

针对企业的个性化流程定制需求, 对业务流程的可变性进行管理也是一个重要的研究课题。Rosemann 提出了可配置流程模型的概念^[98]。一个可配置的业务流程模型代表着一个业务流程家族, 可以看作是业务流程在某一个领域主题内的知识聚集体。昆士兰科技大学 BPM 组的 Marcello 等在业务流程的可变性管理方面做了较为系统的研究工作。Marcello 提出了一种业务流程自动合并算法^[99], 首先, 将流程家族中的业务流程模型抽象成为有向图, 然后进行概念对齐, 最后通过一个合并算法生成可配置的流程图。文献 [100] 提出了一种基于调查问卷的可配置业务流程模型的配置方法。Wil 等提出了针对可配置业务流程模型配置的验证方法^[101]。

有效的业务流程管理必须能够保证业务流程在运行过程中适应环境的变化, 实现这种变化的能力称为业务流程的适应性。知识工程技术在流程的适应性管理中也得到了应用。Buccharone 等提出了一种情境感知的业务流程适应性管理框架^[102]。经检测, 当流程的某个实例需要被适应性演化时, 这个框架就会自动分析一些相关的流程实例变体的运行轨迹, 并查找与这些运行轨迹对应的可以重现的适应性需求。然后, 与这个当前需要适应性演化的业务流程所处情境进行匹配, 从而可对相关流程实例变体的运行轨迹进行排序并加以复用, 最终对当前业务流程进行适应性演化。在这个框架中, 流程变体的实际运行轨迹及对应的业务情境作为知识被存储并加以复用, 以支持业务流程的适应性管理。

2.4 基于知识的服务质量管理与预测

服务质量管理与预测广泛应用于服务发现、服务推荐以及服务组合等多个环节, 是保障用户质量需求的重要手段。服务质量管理主要是指如何根据用户的质量需求选择合适的服务, 当服务在运行中偏离预期的质量要求时又如何进行替换; 服务质量预测是指对候选服务缺失的服务质量数据进行预测, 为服务质量管理提供完备而准确的数据基础。因此, 服务质量的管理与预测密不可分, 服务质量预测是服务质量管理的重要基础。

在服务质量管理方面, 主要使用的有整数规划方法^[103,104]、混合最优方法^[105]和遗传算法^[106]等。采用整数规划方法可以将全局服务质量约束问题映射为多维背包问题, 然后用整数规划方法寻找服务的最优组合^[103,104]。采用混合最优方法可以将全局服务质量约束映射为质量水平, 然后将获得的质量水平用于局部最优选择中的局部约束, 从而快

速地找到满足全局约束的组合^[105]。采用遗传算法可以将各种服务的组合情况看作遗传基因，并基于各质量属性设计适应度函数，适应度函数反映了服务的总体质量，通过不断迭代最终求解出最优的服务组合^[106]。这些方法大都将服务质量各评价指标加权为一个单目标函数，并基于这个单目标函数解决全局服务质量最优问题，但不能根据用户服务质量要求中的多个条件，进行多目标优化。当服务数量较多时，这三类方法都存在时间复杂度高的缺点。另外，已有的服务质量管理方法绝大部分都是单纯从服务角度进行，当前也出现了少量从用户感知角度出发进行服务质量管理的方法，这些方法注重服务作用于用户感知之后产生的用户体验（Quality of Experience, QoE），并以 QoE 作为服务质量的最终评价标准^[107]。面向 QoE 的服务质量管理方法研究服务质量各评价指标对于用户感知的作用机理，并通过这种作用机理对服务质量进行评价，因而理论上可以获得准确的、用户角度感受到的服务质量。但这种方法仍处于起步阶段。

在服务质量预测方面，主要使用的是协同过滤（Collaborative Filtering, CF）方法。CF 方法分为两种，一种是基于近邻的 CF，另一种是基于模型的 CF。基于近邻的 CF 主要分两个步骤：发现相似的用户或服务；基于相似用户或服务进行 QoS 缺失数据的预测。基于近邻的 CF 大都以各自的方法对上述两个步骤进行改进，并致力于提高预测的准确度^[108-110]。基于模型的 CF 是近几年在机器学习知识发现领域发展起来的新方法，主要使用矩阵分解（Matrix Factorization, MF）技术，针对矩阵分解中的优化目标以及避免过拟合问题进行改进^[111,112]。基于矩阵分解的线性偏差因子（Linear Bias Factor, LBF）预测模型也是基于模型的 CF 常用的方法，相关方法在矩阵分解的基础上又增加了若干因子，分别表示系统、用户以及服务所固有的、稳定的、不受外界干扰的特征^[113,114]。随着网络技术的发展，在服务质量预测时需要考虑更多的辅助信息，因此又出现了时间感知的、位置感知的服务质量预测方法^[115-116]。基于近邻的 CF 主要受到传统推荐系统（如音乐、电影以及书籍推荐）预测方法的影响，而没有注意传统推荐系统预测的都是主观数据，而服务质量的各指标参数是客观数据，仅简单借鉴 CF 思想而不做本质改进，可能会导致预测结果产生较大误差。基于模型的 CF 和 LBF 方法处理的都是“用户 - 服务”二维矩阵，不能包含时间、位置等信息；而现有的考虑时间、位置等因素的预测方法，又是针对单个因素的特点进行的，不能将所有的因素统一地考虑在一起，以准确反映服务质量数据的多维结构。

3 国内研究进展

3.1 基于知识的服务发现与推荐

(1) 服务发现

如图 5 所示，国内主要的服务发现方法目录也可按照类似于图 2 的结构进行分

类^[117-123]。针对支持非描述逻辑的方法，国内对于服务发现方法在集中式和分布式两类中的研究都有所进展。文献[119]设计了一个基于本体论的面向服务的P2P架构，用于分布式制造环境中跨虚拟企业的基于语义的制造服务发现。在支持描述逻辑和混合型的两类情况中（尤其是混合型），主要研究成果集中于分布式方法。文献[120]提出一个基于P2P、名为Chord4S的分布式服务发现方法，Chord4S利用数据分布以及热门Chord的查找功能，用分布式方法分发和搜索服务。

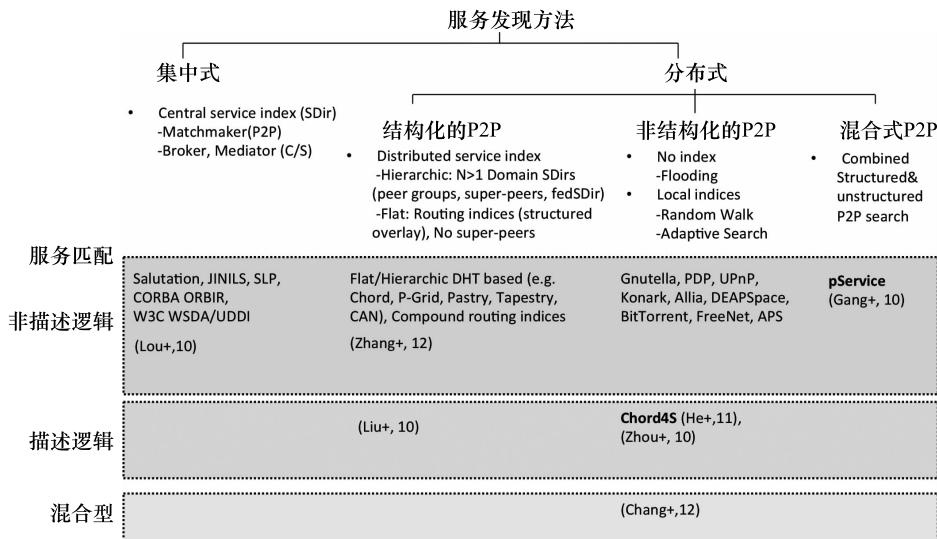


图5 国内现有主要服务发现方法

(2) 服务推荐

如图6所示，国内主要的服务推荐方法目录也可按照类似于图3的结构进行分类^[124-146]。从垂直方向上看，仅考虑服务功能性信息时，主要研究集中在基于内容的服务推荐。文献[124]根据用户的历史信息提取用户的功能兴趣和QoS偏好，通过优化新的相似度指标为用户推荐Top-K个服务。使用服务非功能性的研究涉及了三类推荐方法，取得了较好的研究进展。文献[133]设计了一个Web服务推荐系统WSRec。该系统包含user-contribution机制用于QoS收集，并利用混合协同过滤算法预测服务的QoS值。在同时使用服务的功能性信息和非功能性信息方面，三类方法的研究都较为热门。文献[139]提出了情境感知协同过滤方法来实现服务推荐。该方法主要利用时间、位置以及用户兴趣三种上下文信息。

3.2 基于知识的服务组合

国内在基于知识的服务组合问题上的研究进展也符合上述分类，代表性研究结果简要归为以下类别，具体可参阅相关研究综述。

1) 云制造环境下的服务组合^[148-149]：将服务组合扩展到传统的制造领域，通过对制

造资源的虚拟化形成云制造环境，探索该环境下的服务组合方法。这里的知识主要体现为制造领域的领域知识。

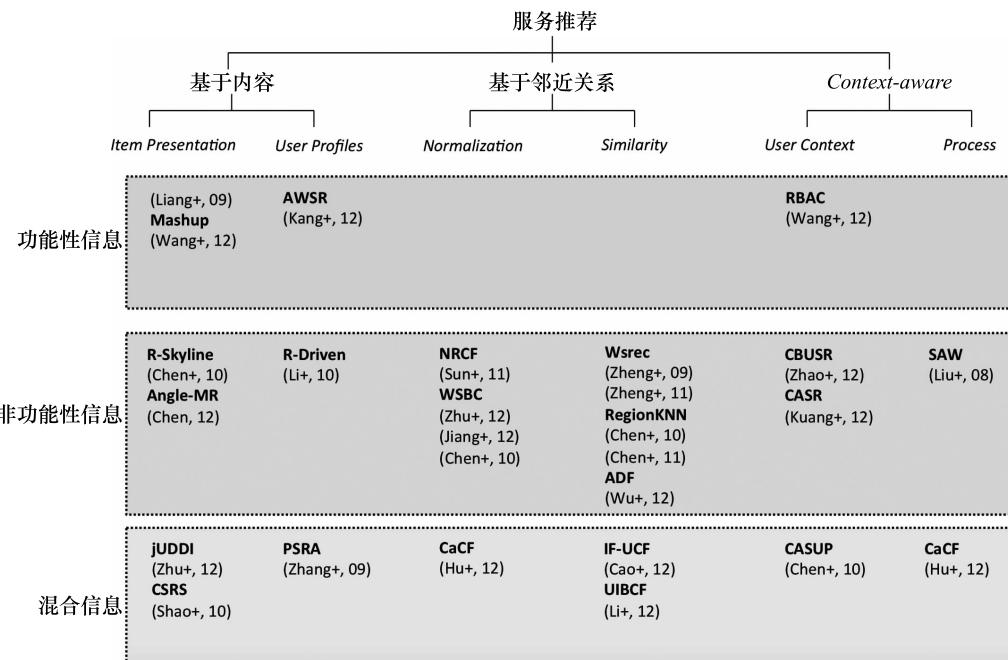


图 6 国内现有主要服务推荐方法

2) 语义服务组合中的知识表示与推理^[67, 70, 150]: 将描述逻辑、时序逻辑、体系结构描述语言、认知模型、进程代数、Petri 网等知识表示语言/形式化方法应用到服务组合中，利用其各有侧重的推理能力对服务组合过程和结果进行各种验证（行为兼容性分析、性能分析等）。

3) 上下文感知的服务组合^[151]: 将客户、候选服务、组合服务所处的上下文环境中所蕴含的信息和知识充分应用到服务组合过程中，提升组合服务的自适应能力。

4) 对传统服务组合算法的改进和提升^[59, 64, 68, 74, 152-153]: 针对 QoS 感知的组合算法，将蚁群算法、遗传算法、粒子群算法、人工蜂群算法等智能算法与服务领域知识紧密结合；针对基于 AI 规划的组合算法，探索新型的启发式策略和并行搜索策略，得到具有更快收敛速度、更好优化效果的新型服务组合算法。

5) 基于服务社交关系的组合策略^[80, 75, 76154]: 不再将各候选服务看作独立的实体，关注服务之间的潜在社交关系，利用此类关系探索式地构造服务网络，可直接或通过定制满足客户需求。

6) 多粒度多层次的服务组合方法^[52, 155, 160]: 与服务社交关系类似的思路，将历史上频繁一起使用的服务提前组织起来形成大粒度模式（服务簇），利用历史案例消除组合服务中的缺陷，提升组合效率和质量。

3.3 基于知识的业务流程管理

在国内主要开展基于知识的业务流程管理的研究机构是清华大学、武汉大学、复旦大学、北京航空航天大学、南京大学等高校。从 2011 年开始举办的中国业务流程管理大会 CBPM，更是为国内研究学者提供了一个进行交流的平台。国内研究学者在基于知识的业务流程管理领域的工作主要围绕流程挖掘、流程的可变性管理、流程的适应性管理、业务流程的互操作性管理等方面展开。

(1) 业务流程知识的挖掘

清华大学在业务流程的知识挖掘工作方面取得了一系列成果。文献 [156] 提出了 α # 算法，用来对含有不可见任务（如，路由网关）的事件日志进行挖掘。 α # 算法可看做对 Wil 的 α 算法的改进，客观上扩展了 α 算法的应用场景。由于近年来国际国内提出了不少流程挖掘算法，而缺少一个能被广大学者接受的算法的评价准则，在文献 [157] 中提出了一种针对流程挖掘算法的推荐算法，能够为一个给定的流程集合推荐最佳的流程挖掘算法。另外，复旦大学、南京大学、武汉大学也都在流程的知识挖掘方面取得了一定的成果。文献 [158] 基于领域知识对事件日志进行分类，使用多种现有的挖掘算法对每一类子日志产生一组流程模型，然后，借助遗传算法的优化能力，从中整合得到高质量的业务流程模型。文献 [159] 提出了使用日志内容检查模型结构正确性与使用模型结构检查日志内容完整性的双向检查标准，并提出了一种内容特征与模型结构特征一一对应的新型日志 Token Log，用于流程模型与系统日志的符合性检查，使得检查和判断过程更加清晰，结果更加明确。在流程模型的修补方面，文献 [160] 基于图搜索技术提出了一种可观测行为保持的流程修补方法。首先，抽取出流程模型的可观测行为，并用行为标记模型加以描述；然后，基于图搜索技术在服务网络上进行搜索，重构出符合原流程行为的新的流程模型，从而达到修补的效果。

(2) 知识在业务流程管理中的应用

武汉大学从知识表达的 W5H (Who, Why, How, What, When, Where) 视角，研究提出了针对个性化、多样化复杂需求描述的角色 (Role)、目标 (Goal)、流程 (Process)、服务 (Service) 4 个基本要素及 9 种语义关联关系的 RGPS 元模型框架^[162]，并在本体-RGPS 元模型层上，提出了知识 - 语义的整合刻画理论与方法。为基于知识的服务信息空间描述提供了 RGPS 基本内容标签及其语义化基本矢量的理论基础，进一步拓展研究了以业务流程为了中心的 RGPS 元建模及其模型一致性验证的理论和方法，在基于本体的业务流程管理方面取得了一系列成果。特别是，在基于本体的流程模型互操作性治理和管理方面，自 2007 年开始，武汉大学承担了基于本体-RGPS 的模型互操作性治理和管理的 ISO 国际标准系列的研制，通过抽取出本体、目标、角色、流程、服务等模型的元模型，以及这些元模型之间的关联关系，完成模型的互操作性注册，支持企业对这些异质异构模型互操作性治理和管理以及流程的发现^[161-167]。在流程的可变性管理方面，文献 [168] 提出了一种基于本体的流程配置方法，通过构建流程可变点本体，完

成情境制导的可配置业务流程的自动配置。文献 [169] 利用着色因果网对可配置业务流程建模，对可配置流程模型在配置过程中的业务目标的满足性进行验证与分析。北京航空航天大学在流程的相似性度量方面取得了很好的进展。文献 [170] 给出一种支持不同粒度的流程模型的相似性计算方法，支持流程粒度的组合服务发现与应用。文献 [171] 提出了基于流程结构树的模型元素匹配关系构建技术，并进一步给出了基于树编辑距离的流程模型相似度度量方法。中国移动通信公司结合中国移动通信中的办公自动化流程家族，做出了流程模型碎片化、聚类与合并的实证实验，为流程可变性管理提供了一个不错的工业界案例^[172]。

3.4 基于知识的服务质量管理与预测

在服务质量管理方面，国内相关学者对国际上传统方法的缺点进行了有益的改进。文献 [173] 提出了一种基于全局服务质量约束分解的服务动态选择方法，该方法在服务选择的时间花费和全局服务质量最优之间进行了合理的折中，以得到一个在开放式服务环境下满足用户服务质量要求的时间花费较低的最优解或近似最优解。文献 [174] 提出一种基于多目标遗传算法的服务质量管理方法。通过将全局服务质量约束问题转化为一个多目标服务组合优化问题，利用多目标遗传算法的智能优化原理，同时优化多个目标参数，即在不同的目标之间取均衡，最终产生一组满足约束条件的最优非劣服务聚合流程集，用户可以根据特定需要从中选择最满意的聚合流程。在面向 QoE 的服务质量管理方面，国内的研究还很少。目前国内 QoE 的相关研究基本上是针对多媒体、电话通信等传统业务进行，研究手段基本上是根据特定行业的经验，真正关于用户感知机理方面的研究尚不多见。

在服务质量预测方面，国内相关学者提出了很多基于 CF 的服务质量预测方法。文献 [175] 对于传统 CF 中的相似度挖掘方法进行了改进，传统 CF 方法总是寻找最相似的若干用户或项目（即 Top-k 方法）进行缺失数据预测，此时相似度（即相关系数）为负的相似用户或项目就被排除在外，实际上只要负相似度的绝对值足够大，则仍然属于高相似，在对于缺失值的预测方面和正的相似度具有同样的作用。文献 [176] 首次提出了传统 CF 方法在某些情况下不适用于客观的服务质量数据预测的观点，并根据服务质量数据的特点设计了新的预测方法，对传统的 CF 进行了重大的改进。文献 [177] 提出了面向高维服务质量数据的预测方法，将时间、位置、用户、服务以及其他各种维度进行了统一的考虑，适应于当前融合网络环境下高维服务质量数据的预测。

4 国内外研究进展比较

4.1 基于知识的服务发现与推荐

根据图 2 和图 5，国内外服务发现领域研究进展比较归纳如表 1 所示。服务发现方

面，国内外研究差距主要表现在分布式中结构化和非结构化 P2P 方法上。相比较而言，国内在这两个领域内的服务发现研究较为不足。另外，国内对集中式服务发现的研究不多，尤其是在支持逻辑型和混合型两个方向更是如此。

表 1 国内外服务发现研究进展比较

国内外研究 进展比较	集中式 (Centralized)			分布式 (Decentralized)								
				结构化的 P2P			非结构化的 P2P			混合式的 P2P		
	非逻辑型	逻辑型	混合型	非逻辑型	逻辑型	混合型	非逻辑型	逻辑型	混合型	非逻辑型	逻辑型	混合型
国内服务发现	√	×	×	√	√	×	×	√	√	√	×	×
国际服务发现	√	×	×	√	√	√	√	√	√	√	×	×

根据图 3 和图 6，国内外服务推荐领域研究进展比较归纳如表 2 所示。在服务推荐方面，国内外研究差距主要体现在基于协同过滤的邻近关系的方法上。考虑使用服务功能性信息时，国内在基于邻近关系方法的研究较少，而国际上在该方向上的研究不论是基于正则化还是相似度计算都较为成熟。但在使用服务非功能性以及混合信息方面，国际上在基于邻近关系的推荐方法上的进展较为缓慢。国內在这方面的研究较为成熟，研究成果也相对比较显著，尤其是在使用非功能信息领域内。总体上来说，在服务推荐领域，国内研究涉及的范围较为广泛，国际上主要针对使用服务功能性信息方向进行深入研究。

表 2 国内外服务推荐研究进展比较

国内外研究 进展比较	基于内容 (Content-based)			基于邻近关系 (neighbor-based)			情境感知 (context-aware)		
	功能性	非功能性	混合	功能性	非功能性	混合	功能性	非功能性	混合
国内服务推荐	√	√	√	×	√	√	√	√	√
国际服务推荐	√	√	√	√	×	×	√	√	√

4.2 基于知识的服务组合

总体上看，国内外针对基于知识的服务组合问题的研究框架是相同的，国内的研究与国际上基本保持同步，在某些研究点上甚至领先于国外研究者。

相比而言，国内学者更倾向于在理论层面开展研究，尤其是在服务领域知识的表示、基于知识的组合服务验证、对传统服务组合算法的改进、基于大粒度模式的服务组合等方面，近年来取得了较大的理论突破。

国外学者则更倾向于将理论层面的服务组合算法与各个实际应用领域密切结合，注重对领域知识的整理与积累，使普适性的服务组合算法能够更有效地解决实际应用问题。

4.3 基于知识的业务流程管理

在基于知识的业务流程管理领域，经过多年的发展与追赶，国内的研究紧跟国际研

究的热点，在某些方面已经达到了国际先进水平。例如，清华大学对流程挖掘算法的推荐方法进行了研究^[157]，这为后续的研究开辟了一个新的方向。武汉大学提出的 RGPS 核心技术作为异构流程模型之间的语义互操作性注册与存储的治理和管理技术，已经得到 ISO 的认可，并形成了国际标准系列^[161-164]。

但是与国外的研究相比，国内的研究也存在着诸多不足。首先，相比于国外知识性业务流程管理的研究成果，国内在基于知识的业务流程管理领域的研究重大原创性成果还尚显不足，总体上还处于追赶国外同类研究的阶段，大部分研究工作还集中于对国外同类研究的扩展与改进上面。其次，国内研究还缺乏国内企业的参与，这使得国内科研人员在案例研究、实验数据获取方面非常艰难，而更多依赖国外的应用需求，研究成果转化也有欠缺。

4.4 基于知识的服务质量管理与预测

在服务质量管理方面，在传统的基于服务自身的服务质量管理方法上，国内相关研究处于国际领先水平，主要体现在算法的运行效率更高、针对全局服务质量约束的多目标优化更加合理。在新兴的面向 QoE 的服务质量管理方面，国内相关研究较少，但国外的相关研究也不深入，尚未取得重要的成果。

在服务质量预测方面，国内相关研究处于国际领先水平。国内相关研究针对服务质量数据的特点对传统的 CF 方法进行了重大改进，使得其适用于服务质量的预测；在考虑时间、位置以及服务质量属性等高维信息的预测方面，国内相关研究首次提出了面向高维数据的服务质量预测方法，将各种维度进行了统一、整体的考虑，从而能在任意维度上进行准确、方便的预测。

5 发展趋势与展望

5.1 基于知识的服务发现与推荐

随着移动互联网、物联网等的迅猛发展，仅仅从静态、已知、稳定的传统角度研究服务发现与推荐则显得不够，需要在动态未知不稳定情况下实现对服务质量的预测；在海量服务的情境下优化传统服务发现与推荐的效率，更加准确地发现并组合符合需求的一组服务。当前有关这方面的研究国内外才刚刚起步，蕴藏着巨大的机遇和挑战。

5.2 基于知识的服务组合

近年来，服务组合从传统的互联网服务领域向其他各问题领域扩展，诸如普适计算环境和物联网、传感器网络、云计算^[71,179]、社会计算等领域的研究者逐渐将服务组合问

题看作本领域的一个重要的理论研究问题，极大地拓展了服务组合问题的研究范畴，丰富了可用的领域知识，使服务组合方法更符合特定应用领域的需求，提升了研究成果的可用性。

从研究对象的角度看，服务组合问题的研究对象从最初阶段单纯面向 Web 服务向更广义的服务扩展，尤其是随着虚拟化技术的成熟所导致的“万物皆服务”现象，现实世界当中的各类硬件设备所提供的服务、云计算环境中的基础设施/平台/软件服务、社交网络中的人以及人所构成的群体所承载的智慧服务等各类遍布信息和物理世界的实体/行为/信息/数据均成为服务组合的可用候选对象，极大地提升了组合服务满足客户需求的能力^[49,81]。

在这种趋势下，服务组合问题不再被简化为一个数学层面的组合优化问题或规划问题，研究者对多维度、多层次、多粒度的领域知识进行获取、分析、推理、使用，利用这些知识来提升服务组合算法的效率、优化能力和满足客户需求的能力。新涌现出的服务组合方法也越来越贴近现实中的应用场景，而不再被看作纯粹的理论研究问题。

5.3 基于知识的业务流程管理

在基于知识的业务流程管理领域，如何有效地获取、利用流程知识是十分重要的课题。在如下几个方面的研究已经或即将成为基于知识的业务流程管理研究的热点。

在流程知识的挖掘方面，如何抽取合适日志数据往往要比流程挖掘本身更为艰难。这就需要有更好的清洗、合并、寻找合适事件数据集的方法；需要一个公正的基准测试集用来对越来越多的流程挖掘算法进行一致的评价；另外，由于在实际应用中存在着跨组织的业务流程，如何利用跨组织的事件日志挖掘出跨组织的业务流程模型也是一个需要解决的问题。

知识在业务流程的应用方面，虽然目前已有基于图合并的流程合并算法以及可配置流程模型的挖掘算法，但还缺乏真正意义上的可配置流程自动生成算法并用于工业实践；大部分的流程可变性及适应性方法缺乏工业界严格的实践验证。随着云计算时代的到来，如何将流程的可变性及适应性管理与云计算的多租户管理有效结合，也是一个需要解决的问题。

5.4 基于知识的服务质量管理与预测

基于知识的服务质量管理与预测方面的研究深度和广度在不断增加。在服务质量管理方面，传统的基于服务本身的服务质量管理方法已没有太多研究空间，目前正逐步转入到基于用户感知角度的服务质量管理，即研究服务质量属性在用户感知上的作用机理，并在此角度上进行和用户感受一致的服务质量管理，这是计算机领域和脑神经认知领域的交叉学科。

在服务质量预测方面，预测对象正从传统的平面数据转变为多维的立体数据，其预

测方法深受传统推荐系统中预测方法的影响。如何将已经在传统推荐系统中获得成功的协同过滤方法全面推广到高维服务质量数据将是下一步需要研究的问题。另外，和传统推荐系统已经有较成熟的冷启动解决方案不同，如何在冷启动情况下进行准确的服务质量预测仍是一个难题。

6 结束语

本文从基于知识的服务发现与推荐、基于知识的服务组合、基于知识的业务流程管理，以及基于知识的服务质量管理与预测等几个方面出发，综述了国内外服务计算特别是知识型服务计算的相关研究进展，并在此基础上展望了其未来发展方向及研究趋势，希望能对国内从事服务计算相关研究的科研人员有所启发。

致谢

本报告受中国计算机学会服务计算专委会委托撰写。在报告的建议与撰写过程中，专委会主任陈俊亮院士给予了学术指导并提出了宝贵意见，北京邮电大学苏森教授、浙江大学吴朝晖教授、哈尔滨工业大学徐晓飞教授提出了宝贵的修改意见，武汉大学王健、冯在文、马于涛、浙江大学陈亮等专委会委员和专家也参与了本报告的撰写工作，在此一并表示感谢。

参考文献

- [1] JGray. The Fourth Paradigm: Data-Intensive Scientific Discovery [M]. Washington: Microsoft Research, 2009.
- [2] TheOpen Group. SOA Reference Architecture [S]. The Open Group Technical Standard, 2011.
- [3] KVerman, K Sivashanmugam, A Sheth, A Patil, S Oundhakar, J Miller. METEORS WSDI: A Scalable P2P Infrastructure of Registries for Semantic Publication and Discovery of Web Services [J]. Information Technology and Management, Special Issue on Universal Global Integration, 2005, Vol. 6: 1.
- [4] LLymberopoulos, S Bromuri, K Stathis, S Kafetzoglou, and M Grammatikou. Towards aP2P Discovery Framework for an Argumentative Agent Technology Assisted Grid [C]. Proc. Of the CoreGRID Workshop on Grid Programming Model, Grid and P2P systems Architectures, Grid Systems, Tools, and Environments, 2007.
- [5] RRanjan, L Zhao, X Wu, A Liu, A Quiroz, M Parashar. Peer-to-peer Cloud Provisioning: Service Discovery and Load Balancing [J]. Computer Communications and Networks, 2010.
- [6] KElgazzar, H Hassanein, P Martin. Effective Web Service Discovery in Mobile Environments [C]. 2011 IEEE 36th Conference on Local Computer Networks (LCN), 2011.
- [7] ECaron, F Chuffart, H Haiwu, C Tedeschi. Implementationand Evaluation of A P2P Service Discovery System: Application in a Dynamic Large Scale Computing Infrastructure [C]. 2011 IEEE 11th

- International Conference on Computer and Information Technology (CIT), 2011.
- [8] ECaron, F Chuffart, C Tedeschi. When Self-Stabilization Meets Real Platforms: An Experimental Study Of A Peer-To-Peer Service Discovery System[C]. Future Generation Computer System, 2012.
- [9] PKungas, M Matskin. Semantic Web Service CompositionThrough a P2P-Based Multi-Agent Environment [C]. Proc. of the Fourth International Workshop on Agents and Peer-to-Peer Computing, 2006.
- [10] C Caceres, A Fernandez, H Helin, O Keller, M Klusch. Context-aware Service Coordination for Mobile Users[C]. Proceedings 1st eHealth Conference, 2006.
- [11] Gharzouli, Mohamed, Boufaida, Mahmoud. A Distributed P2P- Based Architecturefor Semantic Web Services Discovery and Composition[C]. 2010 10th Annual International Conference on New Technologies of Distributed Systems (NOTERE), 2010: 315-320.
- [12] Gharzouli, M., Boufaida, M. A Generic P2P Collaborative Strategy for Discovering and Composing Semantic Web Services [C]. Fourth International Conference on Internet and Web Applications and Services, 2009: 449-454, 24-28.
- [13] P Haase, R Siebes, F van Harmelen. Expertise-based Peer selection in Peer-to-Peer Networks. Knowledge and Information Systems[M]. Springer, 2006.
- [14] A Loser, C Tempich, B Quilitz, W T Balks, S Staab, W Nejdl. Searching Dynamic Communities with Personal Indexes[C]. Proceedings of International Semantic Web Conference, 2005.
- [15] F BKashani, C C Shen, C Shahabi. SWPDS: Web Service Peer- to- Peer Discovery Service [C]. Proceedings of Intl. Conference on Internet Computing, 2004.
- [16] M Klusch, K- U Renner. Dynamic Re- Planning of Composite OWL-S Services [C]. Proc. 1st IEEE Workshop on Semantic Web Service Composition, IEEE CS Press, 2006.
- [17] U Basters, M Klusch. RS2D: Fast Adaptive Search for Semantic Web Service in UnstructuredP2P Networks[C]. Proceedings 5th Intl. Semantic Web Conference (ISWC), USA, LNCS, 2006.
- [18] McGuire , Pory L P, Milligan Michael Van, Conn Jason, Graessley Joshua, Prats Augustin , Tucker Brian. Efficient Service Discovery for Peer- to- Peer Networking Devices [P]. United States Patent Application, 2013.
- [19] L Xu, X Jie, A Nick, J, Li, K Wu. Adaptive Service Discoveryon Service- Oriented and Spontaneous Sensor Systems[C]. Open Access Research, University of Derby Online Research Archive, 2012.
- [20] A Rosenfeld, C Goldman, G Kaminka, S Kraus. An Agent Architecture for Hybrid P2P Free-Text Search [C]. Proceedings of 11th Intel Workshop on Cooperative Information Agents, Delft, Springer, LNAI4676, 2007.
- [21] S Pohlsen, C Buschmann, C Werner. Integrating a Decentralized Web Service Discovery System into the Internet Infrastructure[C]. IEEE Sixth European Conference on Web Services, 2008: 13-20.
- [22] S Ryu, K Han, H Jang, and Y. Eom. User Adaptive Recommendation Modelby Using User Clustering Based on Decision Tree [C]. 2010 IEEE 10th International Conference on Computer and Information Technology, 2010.
- [23] L Liu, N Mehandjiev, D Xu. Multi- Criteria Service Recommendation Basedon User Criteria Preferences [C]. Proc. of the Fifth ACM Conference on Recommender Systems, 2011.
- [24] Y Qi. Decision Tree Learning From Incomplete QoS to Bootstrap Service Recommendation[C]. 2012 IEEE 19th International Conference on Web Services (ICWS), 2012.
- [25] S Han, M Hassan, C Yoon, E Huh. EfficientService Recommendation System for Cloud Computing Market

- [C]. Proc. of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human , 2009.
- [26] H Kim, Y Choi. EmoSens: Affective Entity Scoring, A Novel Service Recommendation Framework for Mobile Platform[C]. Proc. of the fifth ACM Conference on Recommender Systems , 2011.
- [27] N Thio, K Shanika. Web Service Recommendation Based on Client-Side Performance Estimation[C]. 18th Australian Software Engineering Conference , 2007.
- [28] N Chan, W Gaaloul, S Tata. Collaborative Filtering Techniquefor Web Service Recommendation Based on User-Operation Combination[C]. Lecture Notes in Computer Science , 2010.
- [29] M G CA Cimino, B Lazzerini, F Marcelloni, G Castellano, A M Fanelli, M A Torsello. A Collaborative Situation-Aware Scheme for Mobile Service Recommendation[C]. 2011 11th International Conference on Intelligent Systems Design and Applications (ISDA) , 2011.
- [30] A Ciaramella, M G C A Cimino, B Lazzerini, F Marcelloni. Situation- Aware Mobile Service Recommendation with Fuzzy Logic and Semantic Web [C]. 9th International Conference on Intelligent Systems Design and Applications (ISDA) , 2009.
- [31] L Liu, F Lecue, N Mehandjiev, LXu. Using Context Similarity for Service Recommendation [C]. International Conference on Semantic Computing , 2010.
- [32] H Xiao, Y Zou, J Ng, L Nigul. An Approachfor Context- Aware Service Discovery and Recommendation [C]. IEEE International Conference on Web Services (ICWS) , 2010.
- [33] M Kuo, L Chen, C Liang. Building and Evaluating a Location-Based Service Recommendation System with a Preference Adjustment Mechanism[C]. Expert Systems with Applications , 2009.
- [34] N N Chan, W Gaaloul, S Tata. Composition Context Matching for Web Service Recommendation [C]. IEEE International Conference on Services Computing , 2011.
- [35] N N Chan, W Gaaloul, S Tata. Context- Based Service Recommendation for Assisting Business Process Design[C]. Lecture Notes in Business Information Processing , 2011.
- [36] H Yu, S R Marganic. Automated Context-Aware Service Selection for Collaborative Systems[C]. Lecture Notes in Computer Science , 2009.
- [37] L Liu, N Mehandjiev, L Xu. Using Contextual Informationfor Service Recommendation [C]. 44th Hawii International Conference on System Sciences (HICSS) , 2011.
- [38] O Kwon, J Kim. Concept Lattices for Visualizing and Generating User Profiles for Context- Aware Service Recommendations[C]. Expert Systems with Applications , 2009.
- [39] J Rao, X Su. A Survey of Automated Web Service Composition Methods[C]. Semantic Web Services and Web Process Composition, Lecture Notes in Computer Science Volume 3387 , 2005 : 43-54.
- [40] S Dustdar, W Schreiner. A Survey On Web Services Composition[J]. International Journal Of Web and Grid Services. 2005 , 1: 1-30.
- [41] Q Z Sheng, X Qiao, Athanasios V Vasilakos, Claudia Szabo, Scott Bourne, Xiaofei Xu. Web Services Composition: A Decade's Overview[J]. Information Sciences , 280 : 218-238.
- [42] P N Bless, D Klabjan, S Y Chang. Automated Knowledge Source Selection and Service Composition[J]. Computational Optimization and Applications , 2012 , 52 : 507-535.
- [43] G De Giacomo, F Patrizi, S Sardina, Automatic Behavior Composition Synthesis [J]. Artificial Intelligence , 2013 : 196 , 106-142 .
- [44] L Chen, N R Shadbolt, C Goble, F Tao, S J Cox, C Puleston, P R Smart. Towards a Knowledge-Based

- Approach to Semantic Service Composition [J]. The Semantic Web- ISWC 2003 , Lecture Notes in Computer Science , 2003 , 2870 : 319-334.
- [45] RKarunamurthy, F Khendek, & R H Glitho. A Novel Architecture for Web Service Composition [J]. Journal of Network and Computer Applications , 35(2) : 787-802 , 2012.
- [46] M A R C O Comerio. Web Service Contracts: Specification, Selection and Composition. Doctoral Dissertation[D]. Phd Thesis, University of Milano-Bicocca , 2009.
- [47] S ETbahriti, C Ghedira, B Medjahed, M Mrissa, Privacy-Enhanced Web Service Composition[J]. IEEE Transactions on Services Computing , 2014 , 7(2) , 210-222.
- [48] PŚwiątek, P Stelmach, A Prusiewicz, K Juszczyszyn, Service Composition in Knowledge- Based SOA Systems[J]. New Generation Computing , 2012 , 30(2-3) : 165-188.
- [49] T G, Stavropoulos, D, Vrakas, I Vlahavas, A Survey of Service Composition in Ambient Intelligence Environments[J]. Artificial Intelligence Review , 2013 , 40(3) : 247-270.
- [50] MMadkour, D El Ghanami, A, Maach, A Hasbi, Context-Aware Service Adaptation: An Approach Based on Fuzzy Sets and Service Composition[J]. Journal of Information Science and Engineering , 2013 , 29 (1) : 1-16.
- [51] BUpadhyaya, RTang, Y Zou. An Approach for Mining Service Composition Patterns From Execution Logs [J]. Journal of Software: Evolution and Process , 2013 , 25(8) : 841-870.
- [52] GLi, L Liao, D Song, et al. A Self-Healing Framework for Qos-Aware Web Service Composition Via Case-Based Reasoning[J]. In Web Technologies and Applications , 2013 : 654-661.
- [53] W,Dou, Zhang, X Liu, J Chen, J HireSome-II: Towards Privacy-Aware Cross-Cloud Service Composition for Big Data Applications[C]. IEEE Transactions on Parallel and Distributed Systems , 2015 , 26 (2) : 455-466.
- [54] Z Liu,Z Jia, X Xue, J Yu, An Reliable Web Service Composition Based on QoS Dynamic Prediction [C]. Soft Computing , 2015 , 19(5) : 1409-1425.
- [55] J Cheng, C Liu, M Zhou, Q Zeng, A Yla- Jaaski, Automatic Composition of Semantic Web Services Based on Fuzzy Predicate Petri Nets[C]. Automation Science and Engineering , IEEE Transactions , 2015 , 12(2) : 680-689.
- [56] M Alrifai, T Risse, W Nejdl, A Hybrid Approach for Efficient Web Service Composition With End-to-End QoS Constraints[C]. ACM Transactions on the Web , 2012 , 6(2).
- [57] M Alrifai, D Skoutas, T Risse, Selecting Skyline Services for QoS-Based Web Service Composition[C]. In Proceedings of the 19th International Conference on World Wide Web , 2010 : 11-20.
- [58] S X Sun, J Zhao, A Decomposition-Based Approach for Service Composition with Global QoS Guarantees [C]. Information Sciences , 2012 : 199 , 138-153.
- [59] L Wang, J Shen, Yong. A Survey on Bio- Inspired Algorithms for Web Service Composition [C]. In Computer Supported Cooperative Work in Design (CSCWD) , 2012 IEEE 16th International Conference , 2012 : 569-574 .
- [60] O Hatzi, D Vrakas, M Nikolaidou, N Bassiliades, D. Anagnostopoulos, I Vlahavas. An Integrated Approach to Automated Semantic Web Service Composition Through Planning[C]. IEEE Transactions on Services Computing , 2012 , 5(3) : 319-332.
- [61] J Peer, Web Service Composition as AI Planning- a Survey [C]. Technical Report, University of St. Gallen. 2005.

- [62] M Kuzu, N K Cicekli, Dynamic Planning Approach to Automated Web Service Composition[C]. Applied Intelligence , 2012 , 36(1) : 1-28.
- [63] Y Yan, M Chen, Y Yang, Anytime QoS Optimization over the PlanGraph for Web Service Composition [C]. In Proceedings of the 27th Annual ACM Symposium on Applied Computing , 2012 : 1968-1975.
- [64] S Deng, B Wu, J Yin, Z Wu, Efficient Planning for Top-K Web Service Composition[C]. Knowledge and Information Systems , 2013 , 36(3) : 579-605.
- [65] S Song, S W Lee, A Goal- Driven Approach for Adaptive Service Composition Using Planning [C]. Mathematical and Computer Modelling , 2013 , 58(1) : 261-273.
- [66] I Paik, W Chen, M N Huhns, A Scalable Architecture for Automatic Service Composition [C]. IEEE Transactions on Services Computing , 2014 , 7(1) : 82-95.
- [67] Tang, X, Jiang, C, Zhou, M. Automatic Web Service Composition Based on Horn Clauses and Petri Nets [C]. Expert Systems with Applications , 2011 , 38(10) : 13024-13031.
- [68] S Deng, L Huang, W, Tan, Z Wu, Top-K Automatic Service Composition: A Parallel Method for Large-Scale Service Sets [C]. IEEE Transactions on Automation Science and Engineering , 2014 , 11 (3) , 891-905.
- [69] Y Charif, N Sabouret, Dynamic Service Composition Enabled by Introspective Agent Coordination[C]. Autonomous Agents and Multi-Agent Systems , 2013 , 26(1) : 54-85.
- [70] X Wang, W Niu, G Li, et al. , Mining Frequent Agent Action Patternsfor Effective Multi- Agent- Based Web Service Composition[C]. In Agents and Data Mining Interaction , 2012 : 211-227.
- [71] J O Gutierrez-Garcia, K M Sim, Agent- Based Cloud Service Composition [C]. Applied Intelligence , 2013 , 38(3) : 436-464.
- [72] A Lomuscio, H Qu, M Solanki, Towards Verifying Contract Regulated Service Composition [C]. Autonomous Agents and Multi-Agent Systems , 2012 , 24(3) : 345-373.
- [73] F Daniel, C Rodríguez, S Roy Chowdhury, et al. Discovery andReuse of Composition Knowledge for Assisted Mashup Development [C]. In Proceedings of The 21st International Conference Companion on World Wide Web. ACM , 2012 : 493-494.
- [74] X Xu, Z Liu. S- ABC- A Service- Oriented Artificial Bee Colony Algorithm for Global Optimal Services Selection in Concurrent Requests Environment[C]. 2014 IEEE International Conference on Web Services (ICWS) , 2014 : 503-509.
- [75] Q Liang, S Chen, Z Feng. Application of Services Relation Tracing to Automated Web Service Composition[C]. Applied Mathematics & Information Sciences , 2013 , 7(SI) : 243-251.
- [76] S Yan, Y Han, J Wang, C Liu, Service Hyperlink for Exploratory Service Composition [C]. In IEEE International Conference on e-Business Engineering , ICEBE 2007 : 581-588.
- [77] K Vladimir, I Budiselić, S Srblijić. Consumerized and Peer- Tutored Service Composition [C]. Expert Systems with Applications , 2015 , 42(3) : 1028-1038.
- [78] A Maaradji, H Hacid, J Daigremont, N Crespi. Towards a social Network Based Approach for Services Composition[C]. 2010 IEEE International Conference on Communications (ICC) , 2010 : 1-5.
- [79] V Cardellini, E Casalicchio, V Grassi, F L Presti. Flow- Based Service Selection for Web Service Composition Supporting Multiple QoS Classes[C]. In Proceedings of the IEEE International Conference on Web Services , 2007 : 743-750.
- [80] Z Wang, N Jing, F Xu. Cost-Effective Service Network Planningfor Mass Customization of Services

- [C]. International Journal of Services Computing (IJSC) , 2014 , 2(4) : 15-30.
- [81] W Wang, S De, R Toenjes, E Reetz, K Moessner. A comprehensive Ontology for Knowledge Representation in the Internet of Things [C]. In Proc. of 2012 IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom) , 2012 : 1793-1798.
- [82] V Beltran, K Arabshian, H Schulzrinne, Ontology-Based User-Defined Rulesand Context-Aware Service Composition System[C]. In The Semantic Web: ESWC 2011 Workshops , 2012 : 139-155.
- [83] X Liu, G Huang, Q Zhao, H Mei, M B Blake, iMashup. A Mashup- Based Framework for Service Composition[C]. Science China Information Sciences , 2014 , 57(1) : 1-20.
- [84] W M P vander Aalst. Business Process Management: A Comprehensive Survey [C]. ISRN Software Engineering , 2013 : 1-37.
- [85] W MP van der Aalst, A J M M Weijters, and L Maruster. Workflow Mining: Discovering Process Models from Event Logs[C]. IEEE Transactions on Knowledge and Data Engineering , 2004 , 16(9) : 1128-1142.
- [86] J C AM Buijs, B F van Dongen, W M P van der Aalst. Mining Configurable Process Models from Collections of Event logs [C]. Proceedings of the 11th International Conference on Business Process Management (BPM 2013) , 2013 , LNCS 8094 : 33-48.
- [87] J Munoz-Gama, J Carmona, W M P van der Aalst. Single- Entry Single- Exit Decomposed Conformance Checking[J]. Information Systems , 2014 , 46 : 102-122 .
- [88] Dirk Fahland, Wil van der Aalst. Model Repair- Aligning Process Models to Reality [J]. Information Systems , 2015 , 47 : 220-243 .
- [89] N R TP van Beest, E Kaldeli, P Bulanov, J C Wortmann, A Lazovik, Automated Runtime Repair of Business Processes[J]. Information Systems , 2014 , 39 : 45-79.
- [90] W M P van der Aalst, A Adriansyah, B van Dongen, Replaying History on Process Models for Conformance Checking and Performance Analysis [J]. WIREs Data Mining and Knowledge Discovery , 2012 , 2(2) : 182-192.
- [91] W M Pvian der Aalst. Process Mining: Discovery Conformance and Enhancement of Business Process[M]. Springer-Verlag, Berlin , 2011.
- [92] A Rozinat, R S Mans, M Song, W M P van der Aalst. Discovering Colored Petri Nets from Event Logs [J]. International Journal of Software Tools for Technology Transfer , 2008 , 10(1) : 57-74.
- [93] D Martin, M Burstein, J Hobbs, et al. OWL-S: Semantic Markup for Web Services[J/OL]. <http://www.ai.sri.com/daml/services/owl-s/1.2/overview/>.
- [94] O Thomas, M Fellmann, Semantic EPC: Enhancing Process Modeling Using Ontology Languages, Semantic Business Process and Product Lifecycle Management [J]. CEUR Workshop Proceeding of the Workshop SBPM 2007.
- [95] M Hepp, F Leymann, J Domingue, et al. Semantic Business Process Management: A Vision Towards Using Semantic Web Services for Business Process Management[C]. IEEE International Conference on e-Business Engineering (ICEBE 2005) , 2005 : 535-540.
- [96] M Dumas, L Garcia- Banuelos, M La Rosa, et al. Fast Detection of Exact Clones in Business Process Model Repository[J]. Information Systems , 2013 , 38(4) : 619-633.
- [97] H Eberle, T Unger, F Leymann, Process Fragments[C]. OTM 2009 , Part I, LNCS 5870 : 398-405.
- [98] M Rosemann, W M P van der Aalst. A Configurable Reference Modeling Language [C]. Information Systems , 2007 , 32(1) : 1-23.

- [99] M La Rosa, M Dumas, R Uba, R Dijkman, Busienss Process Model Merging: An Approach to Business Process Consolidation[J]. ACM Transaction on Software Engineering and Methodology , 2013, 22 , No. 2.
- [100] M LaRosa, W M P van der Aalst, M Dumas, et al. Questionnaire-Based Variability Modeling for System Configuration[J]. Software and Systems Modeling , 2008, 8(2) : 251-274.
- [101] W M Pvan der Aalst, M Dumas, Gottschalk, Florian, et al, . Preserving Correctness During Business Process Model Configuration[J]. Formal Aspects of Computing , 2008, 22(3-4) : 459-482.
- [102] AntonioBuccharone, Annapaola Marconi, Marco Pistore, et al, . A Context- Aware Framework For Business Process Evolution[C]. The 15th IEEE International Enterprise Distributed Object Computing Conference Workshops , 2011: 146-154.
- [103] L Zeng, Benatallah B, Ngu, A H H, Dumas M, Kalagnanam J, Chang H QoS-Aware Middleware for Web Services Composition[J]. IEEE Transactions on Software Engineering , 2004, 30(5) : 311-327.
- [104] L Zeng, B Benatallah, M Dumas, J Kalagnanam, Q Z Sheng. Quality-Driven Web Services Composition [C]. Proceedings of 12th International Conference on World Wide Web (WWW). New York: ACM , 2003 : 411-421.
- [105] M Alrifai, T Risse, Combining Global Optimization With Local Selectionfor Efficient QoS-Aware Service Composition[C]. Proc. of the 18th Int'l Conf. on World Wide Web. 2009.
- [106] G Canfora, M Di Penta, R Esposito, et al. AnApproach for QoS- Aware Service Composition Based On Genetic Algorithms, Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation. ACM , 2005 : 1069-1075.
- [107] F Lalanne, A Cavalli, S Maag. Quality of Experienceas a Selection Criterion for Web Services[C]. Signal Image Technology and Internet Based Systems (SITIS) , 2012 Eighth International Conference on. IEEE , 2012: 519-526.
- [108] Z Zheng, H Ma, M R Lyu, et al. QoS- Aware Web Service Recommendation by Collaborative Filtering [J]. Services Computing, IEEE Transactions , 2011 , 4(2) : 140-152.
- [109] K Karta. An Investigationon Personalized Collaborative Filtering for Web Service Selection[M]. Honours Programme Thesis, University of Western Australia, Brisbane, 2005.
- [110] H Sun, Z Zheng, J Chen, et al. Nrcf: A Novel Collaborative Filtering Methodfor Service Recommendation[C]. 2011 IEEE International Conference on Web Services. IEEE , 2011 : 702-703.
- [111] W Lo, J Yin, S Deng, et al. An Extended Matrix Factorization Approachfor QoS Prediction in Service Selection[C]. Proceedings of the 9th International Conference on Services Computing. IEEE , 2012: 162-169.
- [112] Z Zheng, H Ma, M R Lyu, et al. Collaborative Web Service QoS Prediction Via Neighborhood Integrated Matrix Factorization[J]. Services Computing, IEEE Transactions , 2013 , 6(3) : 289-299.
- [113] D Yu, Y Liu, Y Xu, et al. Personalized QoS Prediction for Web Services Using Latent Factor Models [J]. Proceedings of International Conference on Services Computing (SCC). IEEE , 2014: 107-114.
- [114] X Luo, M Zhou, Y. Xia, et al. Predicting Web Service QoS Via Matrix-Factorization-Based Collaborative Filtering Under Non- Negativity Constraint [J]. Proceedings of the 23rd Wireless and Optical Communication Conference (WOCC). IEEE , 2014: 1-6.
- [115] A Amin, A Colman, L Grunske. An Approachto Forecasting QoS Attributes of Web Services Based on ARIMA and GARCH Models[C]. Proceedings of the 19th International Conference on Web Services. IEEE , 2012: 74-81.

- [116] M Tang, Y Jiang, J Liu, et al. Location- Aware Collaborative Filtering For Qos- Based Service Recommendation[C]. Proceedings of the 19th International Conference on Web Services. IEEE, 2012: 202-209.
- [117] Y Lou, X Han, Y Wang, S Song. Research on a Web Services Discovery Model Framework [M]. Computational Intelligence and Software Engineering (CiSE), 2010.
- [118] Z Liu, Y Liu, Y He. A Two- Layered P2P Modelfor Semantic Service Discovery [C]. 2010 4th International Conference on New Trends in Information Science and Service Science (NISS), 2010.
- [119] W Zhang, S Zhang, M Cai, Y Liu. A Reputation-Based Peer-to-Peer Architecture for Semantic Service Discovery in Distributed Manufacturing Environments [J]. Concurrent Engineering: Research and Applications, 2012.
- [120] Q He, J Yan, Y Yang, R Kowalczyk, H Jin. A Decentralized Service Discovery Approach on Peer-to-Peer Network[C]. IEEE Transactions on Services Computing, 2011.
- [121] H Chao, J Chen, C Lai. An Efficient Service Discovery Systemfor Dual-Stack Cloud File Service[J]. IEEE Systems Journal, 2012.
- [122] J Zhou, Z Shi. Unstructured P2P-Enabled Service Discovery in the Cloud Environment[J]. Advances in Informantion and Communication Technology, 2010.
- [123] Gang Zhou, Jianjun Yu. pService: Towards Similarity Search on Peer- to-Peer Web Services Discovery [C]. First International Conference on Advances in P2P Systems, 2009: 111-115.
- [124] G Kang, J Liu, M Tang, X Liu, B Cao, Y Xu. AWSR: Active Web Service Recommendation Basedon Usage History[C]. 2012 IEEE 19th International Conference on Web Services (ICWS), 2012.
- [125] L Chen, J Wu, S Deng, Y. Li. Service Recommendation: Similarity-Based Representative Skyline[C]. 6th IEEE World Congress on Services (SERVICES), 2010.
- [126] L Chen, L Kuang, J Wu. MapReduce Based Skyline Services Selectionfor QoS-Aware Composition[C]. International Workshop on High Performance Data Intensive Computing, in conjunction with IPDPS 2012 , 2012.
- [127] G Wang, J Liu, B Cao, M Tang. Mashup Service Classification and Recommendation Based on Similarity Computing[C]. 2012 Second International Conference on Cloud and Green Computing (CGC), 2012.
- [128] A Liu, Q Li, L Huang, S Wen, C Tang, M Xiao. Reputation-Driven Recommendationof Services with Uncertain QoS[C]. 2010 IEEE Asia-Pacific Services Computing Conference (APSCC), 2010.
- [129] J Zhu, Y Kang, Z Zheng, M R Lyu. A Clustering- Based QoS Prediction Approach for Web Service Recommendation[C]. 2012 15th IEEE International Symposium on Object/Component/Service- Oriented Real-Time Distributed Computing Workshops (ISORCW), 2012.
- [130] Z Shao, Z Chen, X Huang. A Mobile Service Recommendation System Using Multi-Criteria Ratings[J]. International Journal of Interdisciplinary Telecommunications and Networking, 2010.
- [131] L Zhang, X Meng, J Chen, K Duan, Y Peng. Personalized Service Recommendation Algorithm. 2nd IEEE International Conference on Computer Science and Information Technology, 2009.
- [132] M Tang, Y Jiang, J Liu, X Liu. Location- Aware Collaborative Filtering for QoS- Based Service Recommendation[C]. IEEE 19th International Conference on Web Services (ICWS), 2012.
- [133] Z Zheng, H Ma, M R Lyu. WSRec: A Collaborative Filtering Based Web Service Recommender System [C]. 7th IEEE International Conference on Web Services (ICWS), 2009.
- [134] Z Zheng, H Ma, M R Lyu, I King. QoS-Aware Web Service Recommendation by Collaborative Filtering

- [C]. IEEE Transactions on Services Computing, 2011.
- [135] X Chen, X Liu, Z Huang, H Sun. RegionKNN: A Scalable Hybrid Collaborative Filtering Algorithmfor Personalized Web Service Recommendation [C]. 9th IEEE International Conference on Web Services (ICWS) , 2011.
- [136] X Chen, Z Zheng, X Liu, Z Huang, H Sun. Personalized QoS-Aware Web Service Recommendationand Visualization[C]. IEEE Transactions on Service Computing, 2011.
- [137] J Wu, L Chen, Y Feng, Z Zheng, M R Lyu, Z Wu. Predicting QoS for Service Selection by Neighborhood-Based Collaborative Filtering[C]. IEEE Transactions on System, Man, and Cybernetics, Part A, 2012.
- [138] H Sun, Z Zheng, J Chen, M R Lyu. NRCF: a Novle Collaborative Filtering Method for Service Recommendation[C]. IEEE International Conference on Web Services (ICWS) , 2011.
- [139] R Hu, W Dou, J Liu. A Context- Aware Collaborative Filtering Approachfor Service Recommendation [C]. International Conference on Cloud and Service Computing (CSC) , 2012.
- [140] L Li, S Pan, X Huang. A Kindof Web Service Recommendation Method Based on Improved Hybrid Collaborative Filtering[C]. IEEE 11th International Conference on Cognitive Informatics & Cognitive Computing (ICCI * CC) , 2012.
- [141] J Cao, Z Wu, Y Wang, Y Zhuang. Hybrid Collaborative Filtering Algorithm for Bidirectional Web Service Recommendation[C]. Knowledge and Information Systems, 2012.
- [142] J Wang, C Zeng, C He, L Hong, L Zhou, R K Wong, J Tian. Context- Aware Role Mining for Mobile Service Recommendation[C]. 27th Annual ACM symposium on Applied Computing (SAC) , 2012.
- [143] L Kuang, Y Xia, Y Mao. Personalized Services Recommendation Based On Context- Aware QoS Prediction[C]. IEEE 19th International Conference on Web Services (ICWS) , 2012.
- [144] D Liu, X Meng, J Chen. A Frameworkfor Context-Aware Service Recommendation[C]. 10th International Conference on Advanced Communication Technology , 2008.
- [145] Z Zhao, N Xu, H Zhan. Context Based Universal Service Recommendation Algorithm [C]. Communications in Computer and information Science, 2012.
- [146] Z Chen, Z Shao, Z Xie, X Huang. An Attribute- Based Schemefor Service Recommendation Using Association Rules and Ant Colony Algorithm[C]. Wireless Telecommunications Symposium, 2010.
- [147] 邓水光, 黄龙涛, 尹建伟, 李莹, 吴健. Web 服务组合技术框架及其研究进展[J]. 计算机集成制造系统, 2011, 17(2) : 404-412.
- [148] 陶飞, 张霖, 郭华, 罗永亮, 任磊. 云制造特征及云服务组合关键问题研究[J]. 计算机集成制造系统, 2011, 17(3) : 477-486.
- [149] 刘卫宁, 刘波, 孙棣华. 面向多任务的制造云服务组合[J]. 计算机集成制造系统, 2013, 19(01) : 199-209.
- [150] 常亮, 刘进, 古天龙, 史忠植. 基于动态描述逻辑的语义 Web 服务组合[J]. 计算机学报, 2013, 36 (12) : 2468-247.
- [151] 唐磊, 淮晓永, 李明树. 一种基于上下文协商的动态服务组合方法[J]. 计算机研究与发展, 2015, 45(11) : 1902-1910.
- [152] 夏亚梅, 程渤, 陈俊亮. 基于改进蚁群算法的服务组合优化[J]. 计算机学报, 2012, 35 (2) : 270-281.
- [153] 李俊, 郑小林, 陈松涛, 陈德人. 一种高效的服务组合优化算法[J]. 中国科学 信息科学 (中文

- 版), 2012, 42(3): 280-289.
- [154] 陈世展, 冯志勇, 王辉. 服务关系及其在面向服务计算中的应用[J]. 计算机学报, 2010(11): 2068-2083.
- [155] 吴洪越, 杜玉越. 一种基于逻辑 Petri 网的 Web 服务簇组合方法[J]. 计算机学报, 2015, 38(1): 204-218.
- [156] L Wen, J Wang, Wil M P van der Aalst, Binqing Huang, Jiaguang Sun. Mimic Process Models with Prime Invisible Tasks[J]. Data Knowl. Eng., 2010, 69(10): 999-1021.
- [157] J Wang, R K Wong, J Ding, et al, Efficient Selection of Process Mining Algorithm, Vol. 6[M]. 2013.
- [158] 杨丽琴, 康国盛, 郭立鹏, 田朝亮, 张亮, 张笑楠, 高翔, 一种适用于多样性环境的业务流程挖掘方法[J]. 软件学报, 2015, 26(3): 550-561.
- [159] 李传艺, 葛季栋, 胡海洋, 胡昊, 骆斌, 一种基于 Token Log 的符合性检查方法, 软件学报[J]. 2015, 26(3): 509-532.
- [160] Z Feng, R Peng, R K Wong, K He, J Wang, S Hu, B Li, QoS-Aware and Multi-Granularity Service Composition[J]. Information Systems Frontiers, 2013, 15: 553-567.
- [161] K He, et al. (Eds.), ISO/IEC 19763-3: Meta-model for ontology registration [S]. <http://metadata-stds.org/19763/index.html#A3>.
- [162] K He, et al. (Eds.), ISO/IEC 19763-5: Meta-model for process model registration [S]. <http://metadata-stds.org/19763/index.html#A5>.
- [163] K He, et al. (Eds.), ISO/IEC 19763-7 Meta-model for service model registration [S]. <http://metadata-stds.org/19763/index.html#A7>.
- [164] K He, et al. (Eds.), ISO/IEC 19763-8 Meta-model for role and goal model registration [S]. <http://metadata-stds.org/19763/index.html#A8>.
- [165] J Wang, N Zhang, C Zeng, Z Li, K He, Towards Services Discovery based on Service Goal Extraction and Recommendation[C]. Proceedings of the 10th IEEE International Conference on Services Computing (SCC 2013), 2013: 65-72.
- [166] 刘建晓, 何克清, 王健, 余敦辉, 冯在文, 宁达, 张秀伟, RGPMS 制导的按需服务组织与推荐方法[J]. 计算机学报, 2013, 36(2): 238-251.
- [167] J Wang, Z Feng, J Zhang, P C K Hung, K He, L Zhang, A Unified RGPMS-Based Approach Supporting Service-Oriented Process Customization[C]. Web Services Foundations, Springer, 2014: 657-682.
- [168] Y Huang, Z Feng, K He, Y Huang, Ontology-Based Configuration for Service-Based Business Process Model[C]. 2013 IEEE Tenth International Conference on Services Computing (SCC 2013), 2013: 296-303.
- [169] 黄贻望, 何克清, 冯在文, 黄颖, 谢芳, 一种基于 RGPMS 着色的 C-net 模型及其应用[J]. 计算机研究与发展, 2014, 51(9): 2030-2045.
- [170] 黄子乘, 怀进鹏, 刘旭东, 李翔, 朱蒋俊, 一个基于流程相似性的自动服务发现框架[J]. 软件学报, 2012, 23(3): 489-503.
- [171] 凌济民, 张莉, 基于过程结构树的过程模型变体匹配技术[J]. 软件学报, 2015, 26(3): 460-473.
- [172] X Gao, Y Chen, Z Ding, et al. Process Model Fragmentization, Clustering and Merging: An Empirical Study[C]. Business Process Management Workshops, 2013: 405-416.
- [173] 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, 22(7): 1426-1439.

- [174] 刘书雷, 刘云翔, 张帆, 唐桂芬, 景宁. 一种服务聚合中 QoS 全局最优服务动态选择算法 [J]. 软件学报, 2007, 18(3): 646-656
- [175] L Shao, J Zhang, Y Wei, et al. PersonalizedQoS Prediction for Web Services via Collaborative Filtering [C]. Proceedings of International Conference on Web Services. IEEE, 2007: 439-446.
- [176] Y Ma, S. Wang, PCK Hung, C-H Hsu, Q Sun, F Yang. A Highly Accurate Prediction Algorithm for Unknown Web Service QoS Value [C]. IEEE Transactions on Services Computing, 2015, PP(99): 1-10. 10. 1109/TSC. 2015. 2407877.
- [177] Y Ma, S Wang, F Yang, R N Chang. Predicting QoS Values via Multi-Dimensional QoS Data for Web Service Recommendations [C]. 2015 IEEE Conference on Web Services (ICWS) Research Track.
- [178] S CGeyik, B K Szymanski, P Zerfos. Robust Dynamic Service Composition in Sensor Networks [C]. IEEE Transactions on Services Computing, 2013, 6(4): 560-572.
- [179] A Jula, E Sundararajan, Z Othman. Cloud Computing Service Composition: A Systematic Literature Review [J]. Expert Systems with Applications, 2014, 41(8): 3809-3824.

作者简介

何克清 武汉大学教授、博士生导师, 日本北海道大学工学博士, 武汉大学软件工程国家重点实验室创始人, 曾长期担任软件工程国家重点实验室主任与学术带头人。目前担任武汉大学软件工程研究所所长, 武汉软件工程学会理事长, ISO/IEC SC32 中国代表, 主持研制完成 ISO/IEC 19763—3, 5, 7, 8, 9 国际标准系列, CCF 服务计算专委会副主任, CCF 杰出会员。研究方向包括软件工程、服务计算等。已发表学术论文 200 多篇、出版专著 4 部。他还担任国际期刊《IEEE Transactions on Service Computing》副主编, 《International Journal of Web Services Research》的编辑。主持国家 973 课题、863 课题、自然科学基金等多项。1990 年获国家颁发的为国家重点实验室建设发展做出了突出贡献的个人“金牛奖”, 2011 年获 ISO 与 IEC 联合颁发的“特别贡献”国际奖, 以第 1 完成人身份获国家科技进步奖二等奖 1 项, 省部级及 CCF 科技进步奖一等奖 5 项、二等奖 2 项等 16 项奖励。



王尚广 副教授、博导。2011 年毕业于北京邮电大学获得博士学位, 2013 年北京交通大学信息与通信工程博士后流动站出站, 2013 年 5 月进入北京邮电大学网络与交换技术国家重点实验室工作。研究方向包括服务计算、云计算、车联网等。目前在 IEEE TSC、IEEE TCC 等 SCI 期刊上发表论文超过 40 篇。主持国家自然科学基金 2 项、北京市自然科学基金 1 项、中国博士后科学基金 1 项, 参与 973 课题、863 课题等多项。2013 年被香港中文大学深圳研究院特聘为副研究员、2015 年当选为国际服务学会中国分会主席。曾担任 IJWSG、IEEE SJ、JOCS 等多个 SCI 期刊的编委、特邀编辑, IOV 2014、SC2



2014 程序委员会主席以及 APSCC 2014 Special Track、IEEE SCC 2015 Application Track 等共同主席。

吴 健 教授/博导，于浙江大学计算机学院获得学士、博士学位，2004 年起在该学院任教至今。浙江大学电子服务研究中心副主任，中国计算机学会青工委委员，中国计算机学会服务计算专委会委员，中国计算机学会计算机应用专委会委员，浙江省 151 人才，科技部重点领域创新团队成员，曾任 YOCSEF 杭州主席，浙江省计算机学会青工委主任。担任 IJSC 杂志编委，PAKDD2013/2014、ICESS2013、ADMA2013 等国际学术会议程序委员会委员，TKDE、KAIS、TSMC、TSC、JWSR 等学术期刊的审稿专家。研究兴趣集中在服务计算、数据挖掘等方面。近年来承担国家科技支撑项目 1 项，国家自然科学基金项目 4 项，浙江省自然科学基金 1 项，863 计划 3 项，浙江省重大科技攻关 1 项。先后在《IEEE Intelligent Systems》、IEEE TKDE、IEEE TSMC、KAIS 等国内外期刊会议发表 SCI/EI 收录论文 90 余篇，所著论文曾获 2008 和 2009 年度中国百篇最具影响国内学术论文称号。2007 年获教育部科技进步一等奖，2008 年获浙江省科技进步一等奖，2009 年获中国商业联合会科学技术特等奖，2010 年获国家科技进步奖二等奖，2014 年获浙江省科技进步一等奖。



王忠杰 1978 年生，工学博士，哈尔滨工业大学计算机科学与技术学院教授，博士生导师，CCF 高级会员，CCF 服务计算专委会委员、软件工程专委会委员。主要研究方向是服务计算、移动与社交服务、软件工程。主持多项国家自然基金项目和 863 项目，在国内外学术期刊与国际会议发表论文 50 余篇。



深度学习与媒体计算

CCF 多媒体专委会

摘要

海量数据的快速增长给多媒体计算带来了深刻挑战。与传统以手工构造为核心的媒体计算模式不同，数据驱动下的深度学习（特征学习）方法成为当前媒体计算主流。本报告重点分析了深度学习在检索排序与标注、多模态检索与语义理解、视频分析与理解等媒体计算方面的最新进展和所面临的挑战，并对未来的发展趋势进行展望。报告总结的方法在深度学习框架下为解决异构鸿沟和语义鸿沟带来了新的思路。

关键词：深度学习，媒体计算，视频分析，视频检索，多模态

Abstract

The increasing large scale data puts forth a great challenge to multimedia computing. Different from traditional multimedia computing which is heavily based on hand- crafted features, deep learning (feature learning) recently achieves noticeable advance in multimedia computing. This report presents the details of multimedia retrieval and annotation, multi- modal semantic understanding as well as the video analysis and understanding. These approaches tend to overcome the heterogeneity gap and semantic gap of multimedia computing in the setting of deep learning framework . This report also outlines several future research challenges in deep multimedia computing.

Keywords: Deep learning, Multimedia computing, Video analysis, Video retrieval, Multi-modal

1 引言

随着互联网用户所创造的内容迅猛增加，从不同渠道涌现的文本、图像和视频等不同类型媒体数据及用户信息（如评论和社区等）更加紧密地混合在一起，以一种新的形式，更为形象综合地表达语义、主题和事件。这样，当前媒体数据呈现如下特点：多种类型媒体数据（文本、图像、视频及属性等）依赖共存，数据来源（各种平台和应用等）广泛丰富，用户交互（个体和群体参与数据产生）史无前例。

Google 前任研究主管 Perter Norvig 博士于 2010 年在《Nature》杂志发表《2020 Visions》文章指出，“今后 10 年（2010 ~ 2020），文本、图像、视频数据、用户交互信息和各种传感器信息将混合在一起，从搜索角度而言，搜索引擎对检索结果进行内容综合而非罗列数据，是谷歌今后面临巨大挑战。”国家科技部于 2011 年 11 月启动了“面向公共安全的跨媒体计算理论与方法”973 项目，对“跨媒体（cross- media）”的表达建模、语义学习、挖掘推理和搜索排序等核心问题进行理论研究。应该说，当前媒体计算

核心要解决两个难点问题，即克服“异构鸿沟”（heterogeneity gap）和“语义鸿沟”（semantic gap）。

传统媒体计算研究主要从手工构建的底层特征出发，利用机器学习方法来填补异构鸿沟和语义鸿沟。合理构建良好手工特征需要大量的工程经验和专业领域知识，通常情况下难以以为每一个特定任务来有针对性地设计所需特征，而是往往依赖于较为通用的特征描述方法（如 SIFT 和 GIST 等），并在这些特征上进一步进行稀疏特征选择和特征降维等处理。与上述依赖手工构建特征的方法不同，深度学习通常对精心设计的多层（根据网络的复杂程度，通常从 5 到 20 层不等，甚至更多）神经网络进行训练来进行特征学习，从而得到区分性更强的特征描述。这些深层神经网络内部有数百万乃至数亿个需要学习的连接权重，一般需要数百万个具有标注信息的数据样本来优化训练。在通过监督学习来训练深度网络时，会根据最小化损失函数得到梯度值，再利用反向传播机制从输出层向输入层逐层计算每层连接权重的梯度，从而更新神经网络中的参数，使得损失函数在训练集上取得最小值。

在深度神经网络中，每一层的神经网络都通过非线性映射对其前一层输出进一步抽象，通过这种逐层抽象的思路，最终学习产生数据的特征描述（即特征学习）。同时，由于深度神经网络的训练是以端到端（end-to-end）方式来实现的，因而最终学习得到了与任务最为相关的特征、去除了与任务无关的信息。比如在分类问题中，良好的特征描述应当对物体位置、旋转或光照等具有鲁棒性。另一方面，现有研究表明，深度神经网络同时具有较强的泛化能力和迁移能力。大量实验表明，在海量数据集上训练得到的深度网络只需要重新设计监督层中的目标函数，通过较小数据集进行监督微调（finetuning）训练，就可在新的数据集和新的应用上取得良好的效果。因此，在某些仅具有较少监督信息的媒体计算应用中也可利用深度学习，这样就显著减少了训练所需的计算资源。

在大规模数据上所进行的实验已经表明：通过深度学习所得到的特征表示在自然语言处理（词向量学习^[1-5]）、知识图谱构建^[6]、图像分类^[7,8]、场景识别^[9]、人体动作检测^[10]和语音识别^[11-13]等领域表现出良好的性能。近年来的各大媒体处理相关的研究型赛事中，深度学习技术也独占鳌头。例如，在大规模视觉识别挑战赛（ILSVRC）中，自 2012 年基于深度学习的方法取得冠军以来^[7]，近年的各大参赛团队的方法主要都基于深度卷积网络，Google 研究组采用改进的卷积网络 GoogLeNet 在 2014 年的比赛中将图像识别准确率提升到了 93.3%^[8]；在微软图像标题生成挑战赛（MS COCO Image Captioning Challenge）中，Google 利用深度网络提取图像特征，并基于具有长短时记忆特点的反馈神经网络（long short-term memory, LSTM）^[14]取得了冠军^[15]；在 THUMOS 2015 竞赛中^[16]，悉尼科技大学 - 卡耐基梅隆大学、微软亚洲研究院和浙江大学获得了视频动作识别任务的前三名，这三支参赛队伍均将传统密集轨迹运动特征与深度神经网络学习特征结合起来进行视频动作识别。

2015 年 5 月 28 日，Nature 在 521 卷、7553 期为人工智能方向开辟了一个专栏，邀请 Yann LeCun、Yoshua Bengio 和 Geoffrey Hinton 等研究人员就深度学习的历史、发展以及

其在文本、图像、视频、语音、生物和疾病研究等各方面的进展和深度学习未来发展方向进行了详细讨论^[17]。

本报告主要介绍深度学习在图像检索排序、多模态语义学习、跨媒体检索与排序以及媒体哈希索引等方面的研究进展，并对深度学习在媒体计算领域今后研究趋势进行讨论。

2 国际研究现状

2.1 图像检索排序与标注

基于文本/元数据的图像检索始于二十世纪七十年代末期。在这一时期，每一幅图像先通过元数据进行描述，然后就可通过预先定义的元数据描述来实现图像检索。然而，当图像数据规模巨大时，这种元数据标注的方式费时费力，出现了较大的局限性，如：缺乏完备的元数据描述标准；元数据描述结果往往具有太多主观色彩，难以反映图像丰富的视觉内容，也就是所谓的“一幅画胜过千言字”（反过来而言，文本本身所承载的语义描述也是很丰富的，也有“一个字胜过千幅画”的说法）。

为了克服文本检索技术的局限性，基于内容的图像检索在二十世纪九十年代应运而生。与通过元数据描述来进行图像检索不同，基于（视觉）内容的图像检索方法往往通过人工构造（hand-crafted）的底层视觉特征（如颜色直方图、SIFT 等）来进行图像相似度计算，从而实现图像检索。本质上而言，基于内容的图像检索方法是基于底层特征相似度检索方法，这种方法难以弥补底层特征和高层语义之间所存在的鸿沟。微软亚洲研究院的芮勇研究员于 1997 年在基于内容的图像检索领域提出了“相关反馈”技术来弥补这一鸿沟。

二十一世纪初期，随着社交网络兴起，用户自产生数据（user-generated data）成为媒体数据产生的基本形式。由于用户会对图像、视频等媒体数据添加标签，因此，如何利用用户标注标签来对图像蕴含语义进行理解，成为媒体研究领域重点。

但是，上述这些方法仍然使用人工构造的底层视觉特征来表达原始数据，其性能受到极大影响。最近，随着深度学习兴起，深度学习方法也被开始广泛使用于图像检索。文献 [18] 提出了一种称为“多尺度无序池化”（Multi-scale Orderless Pooling）的方法，其通过卷积神经网络来提取图像不同尺度下特征，然后在每个尺度上进行无序 VLAD（Vector of Locally Aggregated Descriptors）编码，最后将这些编码结果拼接起来作为图像特征学习结果。文献 [19] 从排序角度提出了一种基于“细粒度深度排序”方法来进行图像检索，其在卷积神经网络最后一层设计了一个排序损失函数来对深层模型进行优化（从而优化特征学习结果），这一排序损失函数由检索样例、相关图像和不相关图像所构

成的三元组组成。排序损失函数要求检索样例与其相关图像之间相似度高于检索样例与其不相关图像之间相似度。

深度卷积神经网络也在图像概念识别上取得了巨大成功。文献 [20] 提出了一种称为“神经编码”的描述符，其将卷积神经网络全连接层之前的滤波结果特征图（feature map）作为特征学习结果，可取得比将全连接层所输出结果作为特征表示更好的实验效果。类似的，文献 [21] 也表明在进行图像检索时，将卷积神经网络中全连接层之前的滤波结果作为特征学习结果会取得更好的效果。具体而言，其从卷积神经网络不同层提取卷积滤波结果，然后利用 VLAD 将这些特征编码成一个向量，最后基于 L2 距离来实现图像检索。

2.2 多模态检索与语义理解

与单一模态的检索不同，多模态检索与语义理解既需要跨越不同模态数据所存在的“异构鸿沟”，也需要弥补底层特征与高层语义之间的“语义鸿沟”。鉴于近几年来深度学习技术在自然语言与计算机视觉等领域取得的进展，国内外的许多工作者开始使用深度学习的方法来提升多模态检索与语义理解的性能。

近一年来，利用深度学习技术实现多模态数据检索的方法主要可以分为基于全局语义和局部语义的深度学习方法。

(1) 基于全局语义的深度学习

基于全局的语义学习方法将每一对完整的多模态数据作为输入来学习，保持了关联性的多模态数据表达。例如，给定“图像 - 描述句子”这样的配对数据，先通过深层模型来学习整幅图像或整个句子的表达，然后再利用 Hing Loss 或最大间隔等目标函数来优化深层模型，将图像和句子投影到公共映射子空间，使得图像和句子两种跨模态数据之间的关联得以保持，这样优化产生的深层模型可实现不同模态数据的表示。这一方面具有代表性的工作是谷歌公司研究人员研发的 NIC (Neural Image Caption)^[15]。NIC 通过最大化给定图像和与之相关文本描述之间的似然估计，来得到一个能够合理评价文本描述与视觉图像之间相关性的映射函数。

(2) 基于局部语义的深度学习。

与前面在深度学习中提取整幅图像或整个句子的表达不同，基于局部的深度学习不是将整幅图像或整个句子作为特征学习单元，而是将视觉对象以及实体单词作为学习单元。具体而言，基于局部的多模态深度学习先提取图像中的视觉对象（或视觉区域）以及描述句子中的实体单词（及其属性），然后对视觉对象和其对应的实体单词（及其属性）进行对齐（alignment），最后通过学习得到将视觉对象和实体单词内嵌到公共子空间的映射函数。这一种方法的代表性研究是斯坦福大学提出的 DeepFE (Deep Fragment Embeddings) 方法^[22]。该方法先对文本描述句子进行处理，得到（语法关系，实体单词 1，实体单词 2）形式的三元组集合；对图像数据使用 Region-CNN 算法定位出得分最高的视觉区域（视觉对象）；用多实例学习算法（Multi Instance Learning, MIL）对文本三

元组和视觉对象进行对齐学习，来挖掘多模态数据之间的相关性。

近年来，组合语义（compositional semantics）这一概念越来越多的出现在跨媒体检索当中。这一概念起源于自然语言处理领域，与侧重于单个单词意义的词汇语义相比，组合语义关注单个单词之间所形成的更丰富语义，例如词组或短语所表达的语义。多数情况下，单个单词仅能承载有限语义，因为单词很少单独出现，而是多存在于词组或句子之中。同样，一幅图像中的视觉对象之间也存在着各种关联关系。因此，与单个单词或单个视觉对象所表达的有限语义不同，组合语义（如词组或近邻视觉对象）更能够表示文本句子或图像中丰富的语义。因此，组合语义学习成为近几年来多模态检索的研究热点。

以文本句子为例，组合语义的表示可分为两类。一类是基于逻辑形式的表示^[23, 25]，如根据语法表达将句子表达成树形结构。另一类是分布式表示，即将词组或句子所形成的组合语义表示为高维空间中的向量^[3]。随着近年来深度学习技术的快速发展，分布式表示逐渐成为组合语义表示的主要手段。目前最前沿的方法主要是通过递归神经网络来建模和学习组合语义表示。文献[23]首先提出了将递归神经网络用于学习组合语义表示。他们提出的方法在之前的使用神经网络学习单词分布式表示^[1]的基础之上拓展到了学习词组的组合语义表示。这一方法使用递归神经网络将一个句子中多个单词的表示传入神经网络以得到多个单词之间的组合表示，再使用分类器评价组合表示是否来自一个词组，如果是词组，则保留这个词组的表示。这些得到的词组的组合语义表示被用于语法分析当中，并取得了显著的效果。

近年来，国外许多公司（如谷歌，微软）与高校（如斯坦福大学，加利福尼亚大学，多伦多大学）研究机构在多模态数据检索与语义理解方面都有相关的工作。

斯坦福大学的研究人员使用 Neural Talk^[25]来进行局部语义学习，进而获得了多模态数据之间的对应关系。具体而言，Neural Talk 使用 RNN (Recurrent Neural Network) 模型来对文本语义进行学习。在 RNN 模型当中，每次输入一个文本单词，就能够得到一个表达当前单词与该单词之前的单词的综合语义的特征向量。Neural Talk 在多模态数据局部语义对齐方面使用了隐马尔科夫模型，并假设两个在文本中相邻的单词所对应的视觉信息应该相同。Socher R 等人^[26]提出了与学习词组表示^[27]不同的方法，其使用递归神经网络为描述图像的句子学习组合语义表示。在学习组合语义表示的同时，得到了文本描述句子和图像的共有内嵌子空间，即图像和文本在统一空间中进行表示，于是文本和图像之间的语义关联可在这个空间中用余弦距离来衡量。

微软研究院使用基于对象的概率预测的方法以及最近邻方法实现了自然语言文本和视觉图像两个模态数据之间的相互检索。基于对象概率预测的方法先利用对象检测的方法找到视觉图像中出现的物体或对象，然后利用概率统计方法找到描述图像的合适的文本信息。该方法使用深度卷积神经网络对每个检测到的视觉对象进行识别，找到能够描述该对象的单词，然后找到能够以最高似然概率覆盖所有识别出的视觉对象单词的文本语句作为对该视觉图像的描述语句。其所使用的深度卷积神经网络模型是利用在大规模图像识别任务中预训练并通过 MIL 方法进行优化训练得到的。在最近邻（Nearest

Neighbor) 检索中^[28]，每幅图像将会通过 GIST 特征与两个深度卷积神经网络训练得到的深度特征进行表达。其中一个深度卷积神经网络是通过大规模图像分类任务训练得到的，另一个是通过基于对象的概率预测的训练方法得到的。给定一幅视觉图像，最近邻方法通过对图像的语义特征表达找出已有图像中与给定图像最相似的图像（欧式距离最接近的图像），并将这些图像的描述文本作为候选的描述文本，然后通过寻找在候选文本中与其他文本的平均相似度最高的文本作为结果文本。

伊利诺伊大学的研究人员使用异构网络映射学习的方法来学习多模态数的语义表达^[29]。该方法将多模态数据表示成为异构网络的形式，其中节点表示多模态数据的每一个样本，边表示多模态数据之间的关系；通过学习映射矩阵将多模态数据映射到共同的语义空间中去。该方法同样使用深度神经网络来学习多模态数据在异构网络中的表达。

2.3 视频分析与理解

在多媒体和计算机视觉领域，视频分析与理解是一个相对比较新的领域。视频数据的分析与理解相比于图像和文本数据具有更高的难度，也具有更强的挑战性。因为视频数据由大量的图像帧组成，数据量大，同时视频图像帧之间有着时序上的语义关联。也就是说，对视频数据的分析不仅要考虑图像中所固有的空间语义，也要考虑视频中所固有的语义时序性，这往往使得视频分析方法对算法和计算资源也有更高要求。近年来随着计算机视觉领域的发展，视频数据的分析也越来越引起了国内外研究者的关注。尤其是由于近年来深度学习方法在图像数据和文本数据上取得的成功，国内外研究者也开始在视频分析上采用深度学习的方法，并取得了一些成效。

为了对视频的内容进行分析，需要对视频有着一个非常好的表达。借助于深度学习框架，表达视频特征时通常从如下几个方面来考虑：

(1) 静态图像特征。考虑到视频是由大量图像帧组成，静态图像特征不考虑视频帧之间的时序关联，直接采用深度学习方法来对图像进行表达，从而得到视频表达。如文献 [35] 采用 ImageNet 预训练的深层网络对视频图像帧进行特征学习，在不同尺度上对网络中间层的特征图 (feature map) 输出进行采样以及编码，显著提升了视频分类精度。这表明通过合适的编码方式，在只考虑静态视频帧信息的情况下也能形成效果非常好的视频表达。

(2) 时序运动特征。对于视频内容的理解离不开对视频中对象运动的分析。法国国家信息与自动化研究所 INRIA 所提出的密集轨迹方法 (improved dense trajectories, iDT)^[31,32] 在多个视频分析任务 (例如动作识别和多媒体事件检测) 中都取得了非常好的效果，被普遍使用在视频分析中。这个方法的思想是对视频中图像子块进行运动追踪，并在所得到运动轨迹上提取 HOG (Histograms of Oriented Gradients)、HOF (Histograms of Optical Flow) 和 MBH (Motion Boundary Histograms) 等特征。最近，一些深度学习方法试图对连续视频帧进行特征学习，如文献 [33] 和 [34] 尝试使用深度学习方法对视频数据中连续视频帧进行处理，但是所得到的特征表示效果并不是很好，这可能是由于这

些方法没能成功捕捉视频中区别性信息。文献 [35] 提出了一种在时间序列上提取深度特征的方法，其使用了文献 [32] 中的轨迹跟踪方法，在卷积神经网络中间层对应位置的特征图输出上来提取特征作为视频特征表达。

(3) 结合使用静态图像特征和时序运动特征的视频表达。Karen Simonyan 和 Andrew Zisserman 在文献 [36] 中提出了一种同时用两个深度神经网络识别视频中动作的方法。这两个网络分别称为空间网络 (Spatial ConvNet) 和时间网络 (Temporal ConvNet)。空间网络的输入是原始视频帧，其对视频中所出现对象进行识别；时间网络的输入是叠加的光流场，其对视频中的运动进行识别。通过结合使用这两个网络，能够形成视频的更好表达。文献 [37] 在同时使用空间网络和时间网络之外，还引入了一个反馈神经网络 RNN (Recurrent Neural Network) 用来编码长时间运动序列，弥补了之前的方法丢失视频中长段运动信息的不足，取得了不错的效果。

生成视频的表达通常采用的都是视觉单词词袋模型 (Bag of visual words)，一般采用基于超级向量思想的 Fisher 向量编码^[38] 和 VLAD (Vector of Locally Aggregated Descriptors) 编码^[39] 将特征描述聚合为一个高维向量。这两种编码方法采用的分别是软分配的混合高斯模型和硬分配的 K 均值模型作为特征编码的词典。基于这两种编码方法的视频表达在分类准确度上优于传统词袋模型表达方法。并且由于使用这两种编码生成的表达在分类任务中可使用线性分类器，使其能够适用于大规模数据集。文献 [40] 提出了一种使用多层 Fisher 向量的编码方法，该方法利用了视频中空间和时间的信息，在视频动作识别上取得了更好的性能。

视频动作识别是视频分析与理解领域中的一个热点问题。自 2013 年起，UCF (University of Central Florida) 开始组织 THUMOS 竞赛^[41]。竞赛的任务是在大规模多类别的视频数据集上进行动作识别和定位。从 2014 年起，竞赛的主要目的转向了在长段未剪辑的视频上检测动作片段。国内外很多研究机构和高校都积极参加这一竞赛，包括微软亚洲研究院、南京大学、香港中文大学、阿姆斯特丹大学和法国国家信息与自动化研究所 INRIA 等。竞赛的两个任务分别是视频动作识别和动作定位。参赛队主要参赛任务还是视频动作识别，很少有参赛队伍参加定位任务。THUMOS 竞赛中主要采用的方法是通过 Fisher 向量编码^[38] 对密集轨迹特征^[31] 进行编码，并辅以其他特征来进一步提升性能。在 THUMOS 2015 竞赛中^[16]，悉尼科技大学 - 卡耐基梅隆大学、微软亚洲研究院和浙江大学分别获得了动作识别任务的前三名。这三个队的算法都是将密集轨迹特征和深度神经网络学习得到的特征描述融合，以得到更好的运动识别效果。第一名悉尼科技大学 - 卡耐基梅隆大学的方案所采用的是用卷积神经网络中间层的输出（隐概念描述子）作为视频帧的表达，并辅以音频特征^[42]。微软亚洲研究院的方案采用了 Facebook 人工智能实验室所提出 C3D 深度网络模型来对视频进行建模，并利用 MFCC 音频特征作为补充^[44]。浙江大学所采用的方案^[24] 是使用文献 [30] 中所提出的深度网络的隐概念描述子作为密集轨迹动作特征的补充，而没有使用其他特征。

自 2011 年起，NIST (National Institute of Standards and Technology) 开始举办 TRECVID 竞赛，吸引了很多研究机构和高校参加，包括美国卡耐基梅隆大学、法国国家

信息与自动化研究所 INRIA、阿姆斯特丹大学、IBM 等。该竞赛主要注重于大规模视频数据集中复杂视频事件的检测。复杂视频事件检测过程涉及多方面的信息，包括图像信息、视频中的文字以及音频信息等。在竞赛中，各参赛队采用的方案基本都是分别提取相应的图像特征（SIFT、Color、CNN）、音频特征（MFCC、ASR）和运动特征（iDT）等，并使用不同的特征融合方式生成最后的视频表达^[45,46]。

当前，视频分析与理解在计算机视觉和多媒体领域将会是下一个热点问题，但是当前的研究还停留在理论探索阶段，距离实用阶段依然还有很长的路要走，还有很多问题需要解决。

1) **算法效率：**算法效率是当前视频分析中面临的一个主要问题。现有的视频分析的主要算法（例如密集轨迹）的计算复杂度都比较高，这使得这些算法的实用性受到很大影响。在最近三年的 THUMOS 竞赛中，基本没有队伍参加动作定位这一任务，其中一个非常重要的原因就是细粒度视频分析所耗用的计算量非常大。当前视频分析算法效率不高，使得完成动作定位任务十分困难。在没有大规模计算资源（例如大规模计算服务器集群）支撑的情况下，动作定位任务所需的计算开销使任务不可能在规定时间内完成的。这需要研究者不断开发出新的更加有效率的算法。

2) **特征模型：**近年来深度学习方法在图像、文本和语音处理上都取得了非常显著的成效。但是在视频分析领域，基于手工特征的视频分析与处理性能仍然优于（基于深度学习方法）特征学习所取得的性能。一个可能原因是视频数据的复杂性，其本身就包含多方面信息（例如静态的图像信息和时序的运动信息），深度学习方法难以捕捉到最具有代表性的视频特征。另一个原因是深度学习方法在训练过程中需要大量有标注的数据进行训练，如对于图像数据来说，ImageNet 数据集就是一个很好的训练数据集。但是对于视频数据来说，目前还缺乏训练视频深度学习方法的大规模数据集。这里存在一个矛盾：一方面对于当前的算法来说，难以在较短时间内处理完大规模的视频数据集；另一方面对于视频数据深度模型的训练又缺少足够规模的已标注数据集。

3) **语义表述：**现在对于视频内容的分析依然是粗粒度，如对整段视频进行分类。更加细粒度的任务动作定位由于计算量增大变得非常困难。对视频内容语义的分析与理解还需要进一步的研究，包括提升算法的效率和细粒度分析的准确性。

3 国内研究进展

3.1 图像检索排序与标注

与国际上深度学习起步步骤相符，国内研究人员也开始在图像检索排序与标注研究中采用深度学习方法。

文献 [47] 研究了深度学习在基于内容的图像检索 (Content Based Image Retrieval) 中这一问题。其在大规模数据预训练得到的卷积神经网络基础上，提出了如下三种图像检索方法：直接用卷积神经网络学习得到的特征表达进行图像搜索；利用卷积神经网络学习得到的特征表达来学习度量函数；定义损失函数来对卷积神经网络进行优化。

深度学习也正将更精细语义或更丰富语义引入不同类型的数据检索，如文献 [48] 提出了“细粒度的图像检索”这一研究，其通过构建层级数据库和定义评价准则来进行特征学习，同时介绍了一个基于细粒度分类图像表示和图像索引基准系统。

3.2 多模态检索与语义理解

国内同样有一些企业与高校在该领域有相关的工作，如百度、华为以及清华大学、浙江大学和中科院计算所等。

百度公司提出了一种 m-RNN 模型^[49]，结合深度卷积神经网络与递归神经网络进行多模态语义内嵌学习。利用深度卷积神经网络学习得到图像特征，然后将该特征输入到递归神经网络的每次预测数据当中。m-RNN 模型使用长短时记忆 (Long- Short Time Memory, LSTM) 层代替一般的 RNN 递归层，并将视觉语义特征直接作为参数输入到每次文本预测中，而不像 NIC 与 Neural Talk 那样视觉特征只在递归神经网络的初始阶段输入一次。

华为公司使用 m-CNN 模型^[50]进行多模态数据的语义理解。m-CNN 由视觉深度卷积网络、基于单词的匹配卷积神经网络和多层次感知器模型构成。基于视觉的卷积网络将每幅图像表达成一个语义特征并作为匹配卷积网络的一项输入。基于单词的匹配卷积网络以每三个相邻的文本单词为一组进行长度为三的一维卷积操作，并且每次卷积都将视觉语义特征作为其中的一项输入。匹配卷积网络叠加多个卷积与 pooling 层，最终得到一个基于多模态数据的语义特征向量。多层次感知器模型以该向量作为输入，产生文本与图像之间匹配度得分。

浙江大学提出了 DTV (Deep Textual Visual correlation learning) 模型^[51]来学习多模态数据之间的关联。在 DTV 模型中，使用 Recursive NN (Recursive Neural Network) 来根据自然语言文本的语法结构学习得到文本描述句子向量表达。对于视觉模态的数据，DTV 使用深度卷积网络学习得到图像向量表达。然后通过矩阵 - 向量相关性分析，将两种不同模态的向量表达映射到一个共同语义子空间，且保持不同模态数据之间的关联度。另一方面，为了进一步学习图像视觉对象和文本实体之间存在的组合语义，文献 [52] 基于排序优化框架提出了多模态深度组合语义学习方法，其不仅保持多模态数据之间的局部相关性，也保持其全局关联性，在多模态检索排序应用中取得了较好效果。

清华大学提出了面向多模态哈希学习的正交性约束深度学习框架，其融合了深度学习的非线性学习能力以及正交性约束的去冗余特性，能学习跨模态信息之间的复杂关联，实现了多模态数据紧凑表示^[53]。

3.3 视频分析与理解

从视频分析与理解的技术上来讲，国内的高校和研究所和国际上的差距并不是非常大，例如中科院深圳研究所在最近两年的 THUMOS 竞赛中都有着不错的表现。微软亚洲研究院和浙江大学作为最近才参赛的队伍，在 THUMOS 2015 中也获得了不错的成绩。另外在 TRECVID 2014 MED 竞赛中，复旦大学、同济大学和北京理工大学等也取得了不错成绩。

4 国内外研究比较

4.1 工业应用

传统的研究中，往往学术界的成果需要一段较长的时间的发展，成熟后才能渐渐进入工业界的实际应用当中。但深度学习作为一个有着极强的工程背景的理论模型，在学术界兴起的同时也被工业界所广泛重视。由于深度学习涉及海量数据的获取和运算需求，在很多领域，工业界相较于学术界有着更大的优势，活跃在多个研究方向的前沿。深度学习在发展理论同时也被广泛应用，如微软的语音助手 Cortana，Google 的 Google Now，在各大搜索引擎所提供的图像检索中更是早已开始采用深度学习技术。

在国外，不仅各大公司，如 Google，Microsoft，Facebook 等的研究院开始进行深度学习的重点研究，甚至在内部成立了深度学习的专门研究机构，如 Google 的 DeepMind 和 Facebook 的 Fackbook AI Research，百度的深度学习研究院等。阿里、腾讯以及华为也都开始设立研究部门对深度学习进行研究。

各大公司纷纷建设起庞大的计算网络以支撑深度学习所需要的巨大计算资源。如百度在 2015 年的图像识别比赛中使用的服务器集群“敏娟”具有 36 台服务器，每台上具有 4 块 Nvidia Tesla K40m GPU，总内存达到 1.7TB 之多，理论上单精度计算能力达到 0.6PFlops。基于这些强大的计算集群，百度可将图像深度学习，从通常 256×256 分辨率提高到 512×512 分辨率，从而提升性能^[54]。

深度学习不仅仅应用在传统的工程问题中，工业界也开始尝试解决传统方法难以解决的问题。如地图应用中的街景服务能够让用户在足不出户的情况下观察到指定地点附近的场景，但由于每个场景往往只具有若干张照片，因而用户在观察时只能看到指定视角所能见到场景。Google 近期提出的 DeepStereo 算法可在给定两张街景图片的情况下生成不同视角的第三张街景图片，这使得 Google Street 应用可以提供更加流畅和自然的街景效果，让用户身临其境般自由行走在虚拟世界中。又如 Google 和百度都在研发的汽车自动驾驶技术，就利用到深度系统来处理传感器中的数据以进行决策控制。

4.2 算法研究

(1) 图像检索排序与标注

图像检索排序与标注是当前国内外的研究重点和热点，拥有广阔的应用前景。在国内，开展图像检索排序与标注的研究单位主要有浙江大学、清华大学、复旦大学、北京大学、华中科技大学、中国科学院计算技术研究所、中国科学院自动化所、微软亚洲研究院以及 IBM 中国研究院等。企业单位主要有百度、腾讯、360 以及搜狗等，并且都已推出了基于内容的图像搜索服务。

国外图像检索排序与标注这一领域的研究单位主要有麻省理工学院、加州大学伯克利分校、美国哥伦比亚大学、美国伊利诺伊大学以及牛津大学等，他们也各自研发出了一些基于内容的图像检索和标注系统。工业界包括微软以及谷歌等，并且都已在搜索引擎中推出了基于内容的图像搜索服务。虽然国外研究者们在图像检索排序与标注的研究上比国内关注的更早，但随着图像检索排序成为国内的研究热点以及更紧密的国际合作的开展，目前国内的研究水平和研究成果与国外差距正在缩小。

(2) 多模态检索与语义理解

在国内，从事多模态语义理解的单位主要有浙江大学、清华大学、复旦大学、西安交通大学、北京邮电大学、北京交通大学、中科院计算所、中科院自动化所、百度深度研究院、华为诺亚方舟研究院和微软亚洲研究院等。相关研究得到了国家重点基础研究发展计划（973 计划）资助。与前几年相比，理论研究和产业应用水平逐年提升。

4.3 开源软件

近年来大量的深度学习工具被开发并以开源形式分享，极大地推动了深度学习研究的发展。代表性的深度神经网络开源工具 Caffe，Torch 和 Theano 等均为国外研究机构或公司支持和发布。由于深度学习中计算的特点，GPU 是目前最合适的计算设备，Nvidia 不仅赞助研究机构 GPU，还提供了进行科学计算的 CUDA 框架以及专为神经网络设计的 cuDNN 库以方便深度学习的训练。2015 年 6 月发布的 cnDNN 3，相较于之前版本提速甚至达到了一倍之多。且基于 Nvidia 和各大开源社区的合作，Caffe，Torch，Theano 都已经完成了基于 cuDNN 3 的优化。

5 未来挑战和展望

虽然目前已有很多深度学习方法应用于媒体计算，但在深度学习中有效利用不同类型媒体数据在不同层次上所具有的耦合特性仍是学术界高度重视的研究问题，正如 Google 公司首席科学家（原微软研究院杰出科学家）John Platt 博士所指：当前智能计算

面临的一大挑战是对强耦合 (strongly-coupled) 输出的整体估计。深度学习的现有进展可喜，但是仍然面临如下挑战：

在线增量深度学习。人脑具有从不断涌现的数据中持续学习而逐步增长经验知识的特有能力，现有生理研究表明：婴幼儿神经细胞在出生后会持续增多且增加轴突、树突和突触等复杂度。Google 于 2015 年 2 月在 Nature 发表了其结合深度学习和强化学习实现具有“pixels-to-actions”能力、超越人类玩家的智能游戏算法研究成果^[55]，说明从数据中不断学习会逐渐提升算法性能。虽然现有若干在线深度学习方法^[56]，但是进一步将不同深层模型学习算法向在线增量学习方向进行拓展，是值得重视的问题。

深度学习的黑盒子问题。目前深度学习仍然在一定程度上是一个黑盒子问题，即根据特定任务来对深层模型本身结构进行优化设计是一个难点问题。如：新加坡国立大学颜水成教授课题组将传统卷积神经网络中线性函数替换为多层网络，提出了 Network in Network 的框架；将深度学习与条件随机场相结合的 Neural CRF 模型^[56]；Google 在 GoogLeNet 中利用了“神经元之间持续重复经验刺激可导致突触传递效能增加”的赫布理论 (Hebbian theory)。因此，如何根据数据本身以及人类认知机理去设计最优的深层网络结构（如网络层数、每一层中隐藏单元数目以及层和层之间的反馈机制等），并且给出深层网络理论分析，尚需理论进一步深入和突破。

深度学习与众包计算结合。深度学习从数据出发，通过端到端方式来学习数据表达，是一种数据驱动的学习方法。尽管深度学习取得了长足进步，但是这一方法受数据噪声影响。如 YAGO 知识库研究者 Gerhard Weikum 教授在 VLDB2014 和 KDD2014 的知识库 Tutorial 中明确指出，要将数据驱动机器学习方法和众包计算方法紧密结合起来，才能实现知识图谱更好地构建。虽然现有基于深度学习框架的弱标签学习和半监督学习从某些侧面利用了众包数据（如利用微软的点击数据集 Clickage 来提升图像检索性能），但是如果将众包中标注数据进行有效利用来提升单纯依赖于数据驱动模式的深度学习方法性能，是值得投入的重点研究。

致谢

浙江大学计算机学院博士生王柱昊、蒋忻洋、宋骏、廖彬兵、宁可在技术进展报告撰写中提供了很大帮助，特此致谢。

参考文献

- [1] BengioY, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3: 1137-1155.
- [2] MikolovT, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space [C]. ICLR, 2013.

- [3] Mikolov, Tomas, et al. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 2013.
- [4] Xu W, Rudnicky A. Can artificial neural networks learn language models [R]. Computer Science Department Technical Report, 2000.
- [5] MikolovT, Deoras A, Povey D, et al. Strategies for training large scale neural network language model [C]. IEEE, 2011, 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) : 011 : 196-201.
- [6] SocherR, Chen D, Manning C, Ng A. Reasoning with neural tensor networks for knowledge base completion[C]. NIPS, 2013 : 926-934.
- [7] KrizhevskyA, Sutskever I, Hinton G. ImageNet classification with deep convolutional neural networks [C]. NIPS, 2012.
- [8] SzegedyC, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going Deeper with Convolutions[J]. arXiv, 2014 : 1409.4842.
- [9] FarabetC, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling[J]. Pattern, IEEE Transactions on Analysis and Machine Intelligence, 2013, 35(8) : 1915-1929.
- [10] TompsonJ J, Jain A, LeCun Y, et al. Joint training of a convolutional network and a graphical model for human pose estimation[C]. Advances in Neural Information Processing Systems. 2014 : 1799-1807.
- [11] Yu D, Deng L. Automatic Speech Recognition-A Deep Learning Approach[M]. Springer, 2014.
- [12] HintonG, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. Signal Processing Magazine, IEEE, 2012, 29(6) : 82-97.
- [13] Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013 : 8614-8618.
- [14] Hochreiter S, Schmidhuber J. Long Short-term memory[J]. Neural Computation, 1997, 9(8) : 1735-1780.
- [15] Vinyals Oriol, Toshev Alexander, Bengio Samy, Erhan Dumitru. Show and Tell: A Neural Image Caption Generator[C]. CVPR, 2015.
- [16] Gorban A, Idrees H, Jiang Y G, Zamir A R, Laptev I, Shah M & Sukthankar R. THUMOS challenge: Action recognition with a large number of classes, 2015.
- [17] Lecun Y, Bengio Y&Hinton G. Deep learning[J]. Nature, 2015 , 521(7553) : 436-444. <http://doi.org/10.1038/nature14539>.
- [18] Gong Y, Wang L, Guo R & Lazebnik S. Multi- scale orderless pooling of deep convolutional activation features. In Computer Vision-ECCV 2014[M]. Springer International, 2014 : pp. 392-407.
- [19] Wang J, Song Y, Leung T, et al. Learning fine- grained image similarity with deep ranking[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014 : 1386-1393.
- [20] Babenko A, Slesarev A, Chigorin A & Lempitsky V. Neural codes for image retrieval. In Computer Vision- ECCV 2014[M]. Springer International, 2014 : pp. 584-599.
- [21] NgJ Y H, Yang F, Davis L S. Exploiting Local Features from Deep Networks for Image Retrieval[J]. arXiv arXiv preprint, 2015 : 1504.05133.
- [22] Karpathy A, Joulin A, Li F. Deep fragment embeddings for bidirectional image sentence mapping[C].

- Advances in neural information processing systems. 2014(a) : 1889-1897.
- [23] Kwiatkowksi T, ZettlemoyerL, Goldwater S, Steedman M. Inducing probabilistic CCG grammars from logical form with higher-order unification [J]. Stroudsburg, PA: Assoc. Comput. Linguist, 2010, In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing: pp. 1223-33.
- [24] Zettlemoyer LS, Collins M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[J]. Arlington, VA: Assoc. Uncertain. Artif. Intell, In Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence: pp. 658-66.
- [25] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[J]. arXiv preprint : 1412.2306, 2014(b).
- [26] Socher R, Karpathy A, Le Q V, et al. Grounded compositional semantics for finding and describing images with sentences [C]. Transactions of the Association for Computational Linguistics, 2014, 2: 207-218.
- [27] Socher R, Manning C D, Ng A Y. Learning continuous phrase representations and syntactic parsing with recursive neural networks[C]. Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop, 2010: 1-9.
- [28] Jacob D, Saurabh G, Ross G, Margaret M, C Lawrence Z. Exploring Nearest Neighbor Approaches for Image Captioning[J]. arXiv, 2015.
- [29] Chang S, Han W, Tang J, et al. Heterogeneous Network Embedding via Deep Architectures[J]. 2015.
- [30] Xu Z, Yang Y & Hauptmann A. G. A discriminative CNN video representation for event detection[J]. arXiv, 2014, arXiv preprint: 1411.4006.
- [31] Wang H, Kläser A, Schmid C & Liu C L. Action recognition by dense trajectories [C]. 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011: 3169-3176.
- [32] Wang H & Schmid C. Action recognition with improved trajectories [C]. 2013 IEEE International Conference on Computer Vision (ICCV), IEEE, 2013: 3551-3558.
- [33] JiS, Xu W, Yang M & Yu K. 3D convolution neural networks for human action recognition [C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2013: 35(1), 221-231.
- [34] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R & Feiffer L. Large-scale video classification with convolutional neural networks [C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2014: 1725-1732.
- [35] Wang L, Qiao Y & Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[J]. arXiv, 2015, arXiv preprint: 1505.04868.
- [36] Simonyan K & Zisserman A. Two-stream convolutional networks for action recognition in videos[C]. In Advances in Neural Information Processing Systems, 2014: pp. 568-576.
- [37] Wu Z, Wang X, Jiang Y. G, Ye H & Xue X. (2015). Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification[J]. arXiv, 2015, arXiv preprint: 1504.01561.
- [38] Sánchez J, Perronnin F, Mensink T & Verbeek J. Image classification with the fisher vector: Theory and practice[J]. International journal of computer vision, 2013, 105(3): 222-245.
- [39] Jégou H, Perronnin F, Douze M, Sanchez J, Perez P & Schmid C. Aggregating local image descriptors into compact codes[C]. IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2012: 34(9) : 1704-1716.

- [40] Peng X, Zou C, Qiao Y & Peng Q. (2014). Action recognition with stacked fisher vectors. In Computer Vision-ECCV 2014[M]. Springer International, 2014: pp. 581-595.
- [41] Thumos, U. C. F. The first international workshop on action recognition with a large number of classes, 2013. In URLhttp://crcv.ucf.edu/ICCV13-Action-Workshop (Vol. 187).
- [42] XuZ, Zhu L, Yang Y, Hauptmann A G. UTS-CMU at THUMOS 2015.
- [43] QiuZ, Li Q, Yao T, Mei T& Rui Y. MSR Asia MSM at THUMOS Challenge 2015.
- [44] NingK & Wu F. ZJUDCD Submission at THUMOS Challenge 2015.
- [45] DouzeM, Oneata D, Paulin M, Leray C, Chesneau N, Potapov D & Harchaoui Z. The INRIA-LIM-VocR and AXES submissions to Trecvid 2014 Multimedia Event Detection[C]. Trecvid, 2014.
- [46] YuS I, Jiang L, Mao Z, Chang X, Du X, Gan C& Hauptmann A. Informedia@ TRECVID 2014 MED and MER. In NIST TRECVID Video Retrieval Evaluation Workshop[C]. Trecvid, 2014.
- [47] WanJ, Wang D, Hoi S C H, Wu P, Zhu J, Zhang Y& Li J. (2014, November). Deep learning for content-based image retrieval: A comprehensive study[C]. ACM, 2014, In Proceedings of the ACM International Conference on Multimedia: pp. 157-166.
- [48] XieL, Wang J, Zhang B & Tian Q. Fine-Grained Image Search[C]. IEEE Transactions on Multimedia, IEEE, 2015, 17(5) : 636-647.
- [49] MaoJ, Xu W, Yang Y, WangJ, Huang Z and Alan Y. Learning like a Child: Fast Novel Visual Concept Learning from Sentence Descriptions of Images[J]. arXiv, 2015 , arXiv preprint: 1504.06692.
- [50] MaL, Lu Z, Shang L, et al. Multimodal Convolutional Neural Networks for Matching Image and Sentence [J]. arXiv, 2015 , arXiv preprint : 1504.06063.
- [51] SongJ, Wang Y, Wu F, et al. Multi-modal retrieval via deep textual- visual correlation learning[C]. ISIDE, International Conference on Intelligence Science and Big Data Engineering, 2015.
- [52] JiangX, Wu F, Li X, Zhao Z, Lu W, Tang S, Zhuang Y. Deep Compositional Cross-modal Learning to Rank via Local-Global Alignment[C]. ACM Multimedia 2015 (Full Paper).
- [53] WangD, Cui P, Ou M, Zhu W. Deep Multimodal Hashing with Orthogonal Regularization. AAAI 2015.
- [54] Wu R, Yan S, Shan Y, et al. Deep image: Scaling up image recognition [J]. arXiv, 2105 , arXiv preprint: 1501.02876.
- [55] Mnih V, Kavukcuoglu K, Silver D, Rusu A, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015: 518: 529-533.
- [56] Zhou G, Sohn K, Lee H. Online Incremental Feature Learning with Denoising Autoencoders [C]. AISTATS, 2012.
- [57] DoT, Artieres T. Neural conditional random fields[C]. AISTATS, 2010: 177-184.

作者简介

吴 飞 博士，浙江大学计算机学院，教授，博士生导师，主要研究方向为多媒体分析与检索、群智计算和机器学习。



朱文武 博士，清华大学计算机系，教授，博士生导师，主要研究方向为三元空间的多媒体计算、媒体大数据计算、多媒体云计算等。



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。



文本自动生成研究进展与趋势

CCF 中文信息技术专委会

摘要

我们期待未来有一天计算机能够像人类一样会写作，能够撰写出高质量的自然语言文本。文本自动生成是实现这一目的的关键技术。按照不同的输入划分，文本自动生成可包括文本到文本的生成、意义到文本的生成、数据到文本的生成以及图像到文本的生成等。上述每项技术均极具挑战性，在自然语言处理与人工智能领域均有相当多的前沿研究，近几年业界也产生了若干具有国际影响力成果与应用。本文对上述前沿技术的国内外研究现状进行了全面总结，并对发展趋势进行了展望。

关键词：自然语言生成，文本到文本的生成，意义到文本的生成，数据到文本的生成，图像到文本的生成

Abstract

We expect that computers can write high-quality natural language texts like human beings in the near future. Automatic text generation is the key technique for achieving this goal. According to different data types of inputs, automatic text generation techniques include text-to-text generation, meaning-to-text generation, data-to-text generation and image-to-text generation. All the above text generation techniques are very challenging, and they are the frontier research topics in the natural language processing and artificial intelligence fields. In recent years, a few internationally influential achievements and applications have been yielded in academia and industry. In this article, we conduct a comprehensive survey of recent advances of automatic text generation at home and abroad. We also discuss the research and development trends.

Keywords: natural language generation, text-to-text generation, meaning-to-text generation, data-to-text generation, image-to-text generation

1 引言

文本自动生成是自然语言处理领域的一个重要研究方向，实现文本自动生成也是人工智能走向成熟的一个重要标志。简单来说，我们期待未来有一天计算机能够像人类一样会写作，能够撰写出高质量的自然语言文本。文本自动生成技术极具应用前景。例如：文本自动生成技术可以应用于智能问答与对话、机器翻译等系统，实现更加智能和自然的人机交互；我们也可以通过文本自动生成系统替代编辑实现新闻的自动撰写与发布，最终将有可能颠覆新闻出版行业；该项技术甚至可以用来帮助学者进行学术论文撰写，

进而改变科研创作模式。

按照不同的输入划分，文本自动生成可包括文本到文本的生成（Text-to-Text Generation）、意义到文本的生成（Meaning- to- Text Generation）、数据到文本的生成（Data- to- Text Generation）以及图像到文本的生成（Image-to-Text Generation）等。上述每项技术均极具挑战性，在自然语言处理与人工智能领域均有相当多的前沿研究，近几年业界已产生了若干具有国际影响力成果与应用。最值得一提的是，美联社自 2014 年 7 月开始已采用新闻写作软件自动撰写新闻稿件，这大大减少了记者的工作量。美国《洛杉矶时报》也有一种用来撰写突发新闻的应用软件。美国已有多家公司能够提供新闻写作软件与服务，比如美国“自动洞察力”公司（Automated Insights）已采用“语言专家”软件撰写了 3 亿篇报道，包括橄榄球、财经报道。这些进展标志着文本自动生成不再属于纸上谈兵的技术，而是已经对人类工作和生活产生了重大影响。

目前国内学界与工业界对文本自动生成技术的重视程度还不够，普遍缺乏对该方向前沿技术与进展的了解。因此，本技术报告将首次对文本自动生成前沿技术进行综合全面的调研、分析与总结，为国内同行提供一个全面了解文本自动生成技术的重要参考。同时，期望学界和工业界一起努力，尽早实现中文文本自动生成系统，抢占中文文本自动生成技术的制高点。

需要指出的是，自然语言处理领域的自然语言生成技术专指从机器可读数据生成自然语言文本的技术，而本文所介绍的文本自动生成技术的范畴则更加广泛，还包括了文本到文本的生成技术，以及图像到文本的生成技术。

2 文本到文本的生成

2.1 国际研究现状

文本到文本的生成技术主要指对给定文本进行变换和处理从而获得新文本的技术，具体说来包括文本摘要（Document Summarization）、句子压缩（Sentence Compression）、句子融合（Sentence Fusion）、文本复述（Paraphrase Generation）等。国际上对上述不同技术均进行了多年的研究，相关研究成果主要发表在自然语言处理相关学术会议与期刊上，例如 ACL、EMNLP、NAACL、COLING、AAAI、IJCAI、SIGIR、INLG、ENLG 等。国际上几个主要的研究单位包括密歇根大学、南加州大学、哥伦比亚大学、北得克萨斯大学、爱丁堡大学等。需要指出的是，机器翻译从某种程度上也可看作是一种从源语言到目标语言的文本生成技术，但由于机器翻译自身是相对独立的一个研究领域，因此本文的内容不再涵盖机器翻译技术。

2.1.1 文本摘要

文本摘要技术通过自动分析给定的文档或文档集，摘取其中的要点信息，最终输出

一篇短小的摘要（通常包含几句话或上百字）。该摘要中的句子可直接出自原文，也可重新撰写。摘要的目的是通过对原文本进行压缩、提炼，为用户提供简明扼要的内容描述。

根据不同的划分标准，文本摘要主要可以分为以下几种不同类型：

根据处理的文档数量，摘要可以分为单文档摘要和多文档摘要。单文档摘要只对单篇文档生成摘要，而多文档摘要则对一个文档集生成摘要。

根据是否提供上下文环境，摘要可以分为主题或查询无关的摘要和主题或查询相关的摘要。主题或查询相关的摘要在给定的某个主题或查询下，能够诠释该主题或回答该查询；而主题或查询无关的摘要则指在不给定主题和查询的情况下对文档或文档集生成的摘要。

根据摘要所采用的方法，摘要可以分为生成式和抽取式。生成式方法通常需要利用自然语言理解技术对文本进行语法、语义分析，对信息进行融合，利用自然语言生成技术生成新的摘要句子。而抽取式方法则相对比较简单，通常利用不同方法对文档结构单元（句子、段落等）进行评价，对每个结构单元赋予一定权重，然后选择最重要的结构单元组成摘要。抽取式方法应用较为广泛，通常采用的结构单元为句子。

根据摘要的应用类型，摘要可以分为标题摘要、传记摘要、电影摘要等。这些摘要通常为满足特定的应用需求，例如传记摘要的目的是为某个人生成一个概括性的描述，通常包含该人的各种属性，例如姓名、性别、地址、出生、兴趣爱好等。用户通过浏览某个人的传记摘要就能对这个人有一个总体的了解。

文档自动摘要的研究在图书馆领域和自然语言处理领域一直都很活跃，最早的应用需求来自于图书馆。图书馆需要为大量文献书籍生成摘要，而人工摘要的方式效率很低，因此急需自动摘要方法取代人工高效地完成文献摘要任务。随着信息检索技术的发展，文档自动摘要在信息检索系统中的重要性越来越大，逐渐成为研究热点之一。文档自动摘要技术的第一篇论文来自 Luhn (1958)^[1]。经过数十年的发展，同时在 DUC[⊖] 与 TAC[⊖] 组织的自动摘要国际评测的推动下，文本摘要技术已经取得长足的进步。值得一提的是，由南加州大学 Chin-Yew Lin 博士（现就职于微软亚洲研究院）开发的摘要质量自动评估工具 ROUGE[⊖] 的广泛使用也是自动摘要技术快速发展的一个推动力。国际上文档自动摘要方面比较著名的几个系统包括 ISI 的 NeATS 系统^[2]、哥伦比亚大学的 NewsBlaster^④ 系统^[3]、密歇根大学的 NewsInEssence^⑤ 系统^[4] 等。2013 年雅虎耗资 3000 万美元收购了一项自动新闻摘要应用 Summly，标志着新闻摘要技术走向成熟。

目前的文本摘要方法主要基于句子抽取，也就是以原文中的句子作为单位进行评估与抽取。这类方法的好处是易于实现，能保证摘要句子具有良好的可读性。该类方法主

⊖ <http://duc.nist.gov/>

⊖ <http://www.nist.gov/tac/>

⊖ <http://www.berouge.com>

④ <http://www1.cs.columbia.edu/nlp/newsblaster/>

⑤ <http://lada.si.umich.edu:8080/clair/nie1/nie.cgi>

要包括两个步骤：一是对文档中的句子进行重要性计算或排序，二是选择重要的句子组合成最终摘要。第一个步骤可采用基于规则的方法，利用句子位置或所包含的线索词来判定句子的重要性；也可采用各种机器学习方法（包括深度学习方法），综合考虑句子的多种特征进行句子重要性的分类、回归或排序，例如 CRF^[5]、HMM^[6]、SVM^[7,8]、RNN^[9]等。第二个步骤则基于上一步结果，需要考虑句子之间的相似性，避免选择重复的句子（如 MMR 算法^[10]），并进一步对所选择的摘要句子进行连贯性排列（如自底向上法^[11]），从而获得最终的摘要。近几年学界进一步提出了基于整数线性规划的方法^[12-14]以及次模函数最大化方法^[15,16]，可以在句子选择的过程中同时考虑句子冗余性。

不同于上述方法，压缩式文本摘要方法则考虑对句子进行压缩，以在较短长度限制下让摘要涵盖更多的内容。最有代表性的做法为同时进行句子选择与句子压缩^[17-19]，能够取得更优的 ROUGE 性能。除了压缩之外，部分工作还利用句子融合等技术来对已有句子进行变换，得到新的摘要句子^[20,21]。

国际上还有部分研究者研究真正意义上的生成式摘要，也就是通过对原文档进行语义理解，将原文档表示为深层语义形式（例如深层语义图），然后分析获得摘要的深层语义表示（例如深层语义子图），最后由摘要的深层语义表示生成摘要文本。最近的一个尝试为基于抽象意义表示（Abstract Meaning Representation, AMR）进行生成式摘要^[22]。这类方法所得到的摘要句子并不是基于原文句子所得，而是利用自然语言生成技术从语义表达直接生成而得。这类方法相对比较复杂，而且由于自然语言理解与自然语言生成本身都没有得到很好的解决，因此目前生成式摘要方法仍属于探索阶段，其性能还不尽如人意。

上述摘要方法均面向新闻摘要，而近年来针对学术文献的摘要越来越受到大家的重视。一方面，可以利用学术文献之间的引用关系以及引文来帮助进行学术文献摘要^[23]；另一方面，对学术文献进行自动综述也是一个很有意思的研究问题^[24]。更多的有关文本摘要技术的内容可参考综述^[25]。

2.1.2 句子压缩与句子融合

句子压缩与句子融合技术一般用于文本摘要系统中，用于生成信息更加紧凑的摘要，获得更好的摘要效果。

句子压缩技术基于一个长句子生成一个短句子，要求该短句保留长句中的重要信息，也就是重要信息基本不损失，同时要求该短句是通顺的。下面给出一个句子压缩的例子：

原句：But they are still continuing to search the area to try and see if there were, in fact, any further shooting incidents.

压缩后的句子：They are continuing to search the area to see if there were any further incidents.

学界尝试了多种方法实现句子压缩，包括从句子中删除词语^[26]，或对句子中的词语进行替换、重排序或插入^[27]。其中，从句子中直接删除词语的做法因其复杂程度较低而

成为主流方法。研究人员提出多种方法用于实现基于词语删除的句子压缩，包括噪声信道模型^[28]、结构化辨别模型^[29]、树到树的转换^[30]、整数线性规划^[31]等。但就总体效果而言，对于大部分句子，删除的词语一般较少，压缩效果体现并不明显。

句子融合技术则是合并两个或多个包含重叠内容的相关句子得到一个句子。根据目的的不同，一类句子融合只保留多个句子中的共同信息，而过滤无关的细节信息（类似于集合运算中的取交集运算），另一类句子融合则只过滤掉多个句子之间的重复内容（类似于集合运算中的取并集运算）。下面给出两个相关的句子以及人工合并后得到的句子：

句子1：In 2003, his nomination to the U. S. Court of Appeals for the District of Columbia sailed through the Senate Judiciary Committee on a 16-3 vote.

句子2：He was nominated to the U. S. Court of Appeals for the District of Columbia Circuit in 1992 by the first President Bush and again by the president in 2001.

合并后的句子（取交集）：He was nominated to the U. S. Court of Appeals for the District of Columbia Circuit.

合并后的句子（取并集）：In 2003, his nomination by the first President Bush, and again by the second Bush in 2001 to the U. S. Court of Appeals for the District of Columbia sailed through the Senate Judiciary Committee on a 16-3 vote.

针对句子融合问题，MIT的Regina Barzilay和哥伦比亚大学的Kathleen McKeown提出一条流水线算法，包括共同信息识别（identification of common information）、融合网格计算（fusion lattice computation）、网格线性化（lattice linearization）三个步骤^[20]。研究人员针对句子融合问题提出的其他代表性方法包括基于结构化辨别学习的方法^[32]、基于整数线性规划的方法^[33]、基于图最短路径的方法^[34]等。

上述研究均面向英文，少数研究者在网上公开了所使用的数据集，但这些数据集的规模相对较小，覆盖面较窄，业界也没有组织过句子压缩或融合相关的评测。近些年，与句子压缩与句子融合技术相关的学术论文比较少见，这与上述因素不无关系。

2.1.3 文本复述

文本复述生成技术通过对给定文本进行改写，生成全新的复述文本，一般要求输出文本与输入文本在表达上有所不同，但所表达的意思基本一样。文本复述生成技术应用相当广泛，例如，在机器翻译系统中可利用文本复述技术对复杂输入文本进行简化从而方便翻译，在信息检索系统中可利用文本复述技术对用户查询进行改写，在儿童教学系统中可利用文本复述技术将难以理解的文本简化为儿童容易理解的文本。

根据实际的需求，通过文本复述技术得到的输出文本与原文本相比，可以只是一两个词发生了改变（如例1），也可以是整段文本面目全非（如例2）。

例1：all the members of -> all members of

例2：He said there will be major cuts in the salaries of high-level civil servants. =>

He claimed to implement huge salary cut to senior civil servants.

简单的文本复述生成可以通过同义词替换来实现，也可以通过人工或自动构建的复述规则来实现^[35]，例如根据变换状语位置的一条规则，可以获得下面句子的简单复述句子：

输入：He booked a single room in Beijing *yesterday*.

输出：*Yesterday*, he booked a single room in Beijing.

为了实现复杂的文本复述，研究人员提出了基于自然语言生成的方法^[36]、基于机器翻译的方法^[37]与基于支点（Pivot）的方法^[38,39]等。基于自然语言生成的方法模拟人类的思维方式，首先对输入句子进行语义理解，获得该句子的语义表示，然后基于得到的语义表示生成新的句子。基于机器翻译的方法则将文本复述生成问题看作是单语言机器翻译问题，从而利用现有机器翻译模型（例如噪声信道模型）来为给定文本生成复述文本。基于支点的方法则将当前语言中的输入文本翻译到另一种语言（支点），然后将翻译得到的文本再次翻译回当前语言。由于每次翻译过程均要求源语言和目标语言中文本的语义保持一致，因此可以预期最后得到的文本在语义上能跟输入文本保持一致。注意支点语言可以只采用一种语言，也可采用多种语言。例如，下面的例子中采用意大利语作为支点语言，通过两次翻译为输入的英文句子生成复述文本：

输入的英文句子：What toxins are English most **hazardous** to **expectant mothers**?

翻译后的意大利文句子：Che tossinesonopiùpericolosealledonneincinte?

再次翻译后的英文句子：What toxins are more **dangerous** to **pregnant women**?

总体而言，现有方法能够为给定文本生成具有较小差别的复述文本，但是难以有效生成高质量的具有很大差别的复述文本，原因在于对于改写甚多的复述文本而言，一方面难以保证其与原文本的语义一致性，另一方面则难以保证该文本的可读性。近几年已经极少在自然语言处理重要会议上看到文本复述生成相关的学术论文，表明针对该问题的研究已经遇到了瓶颈。

需要指出的是，句子简化（Sentence Simplification）可以看作是一类特殊的复述生成问题，其目的是将复杂的长句改写成简单、可读性更好、易于理解的多个短句，方便用户快速阅读。在实现上仍可采用上述各类方法，例如基于单语言机器翻译的方法^[40]、基于树转换的方法^[41]等。针对句子简化问题的很多研究都采用维基百科[⊖]以及对应的简单维基百科[⊖]数据来进行学习和测试。简单维基百科面向的阅读对象为儿童以及正在学习英语的成人，要求作者使用简单的词汇和简短的句子来撰写文章。一篇简单维基百科文章一般对应一篇普通维基百科文章，因此通过这种文本之间的对应关系能够获取大量的有用语料。爱丁堡大学的 Kristian Woodsend 与 Mirella Lapata 则提出基于准同步文法（Quasi-Synchronous Grammar）与整数线性规划模型将普通维基百科文章简化为简单维基百科文章^[42]。

⊖ <http://en.wikipedia.org>

⊖ <http://simple.wikipedia.org>

2.2 国内研究现状

2.2.1 文本摘要

相比机器翻译、自动问答、知识图谱、情感分析等热门领域，文本摘要在国内并没有受到足够的重视。在文本摘要方面从事过研究的单位包括北京大学计算机科学技术研究所、北京大学计算语言所、哈尔滨工业大学信息检索实验室、清华大学智能技术与系统国家重点实验室等。其中，北京大学计算机科学技术研究所在文本摘要方面进行了长期深入的研究，提出了多种基于图排序的自动摘要方法^[43-46]与压缩式摘要方法^[47]，并且探索了跨语言摘要、比较式摘要、演化式摘要等多种新颖的摘要任务^[48-50]。在学术文献摘要方面，则分别提出基于有监督学习和整数线性规划模型的演示幻灯片的自动生成方法^[51]与学术论文相关工作章节的自动生成方法^[52]。

国内早期的基础资源与评测^①举办过单文档摘要的评测任务，但测试集规模比较小，而且没有提供自动化评价工具。2015年CCF中文信息技术专委会组织了NLPCC评测^②，其中包括了面向微博的新闻摘要任务，提供了规模相对较大的样例数据和测试数据，并采用自动评价方法，吸引了多支队伍参加评测，目前这些数据可以公开获得。但上述中文摘要评测任务均针对单文档摘要任务，目前还没有业界认可的中文多文档摘要数据，这在事实上阻碍了中文自动摘要技术的发展。

近些年，市面上出现了一些文本挖掘产品，能够提供文档摘要功能（尤其是单文档摘要），例如方正智思、拓尔思（TRS）、海量科技等公司的产品。百度等搜索引擎都能为检索到的文档提供简单的单文档摘要。这些文档摘要功能均被看作是系统的附属功能，其实现方法均比较简单。由于这些模块均未参加公开评测，因此其性能不得而知。

2.2.2 句子压缩与句子融合

国内有少数单位与学者对句子压缩问题进行了研究，例如北京大学语言计算与互联网挖掘研究室提出基于对偶分解的句子压缩方法^[53]，清华大学智能信息获取研究小组提出基于马尔科夫逻辑网的句子压缩方法^[54]，等等。而对于句子融合问题的研究，国内单位和学者基本没有涉猎。

国内学者的上述研究仍面向英文数据，主要原因在于缺少相关的中文评测数据，而构建一个高质量的中文句子压缩或融合评测数据集并不简单，需要依靠对语言有深刻理解的标注者。

2.2.3 文本复述

国内有少数单位和学者对文本复述生成进行了一些研究，例如哈尔滨工业大学信息

① <http://www.863data.org.cn>。

② http://tcci.ccf.org.cn/conference/2015/pages/page05_evadata.html。

检索中心与微软亚洲研究院、百度等单位合作，提出利用多种资源（包括多种词典、平行语料等在内）改进基于机器翻译的复述生成方法^[55]、利用多种机器翻译引擎的复述生成方法^[56]，以及面向不同应用的复述生成方法^[57]。

上述研究仍面向英文领域，采用英文数据进行评测，而中文复述生成技术则极少有人涉足，这是一件很令人遗憾的事情。

2.3 发展趋势与展望

文本到文本的生成包括多项任务，这些任务之间具有紧密的联系，很多方法也都对不同任务具有通用性。在未来几年，随着深层语义分析技术的发展，研究者可以在研究过程中充分利用深层语义分析结果，此外，深度学习技术的成熟则为我们的研究打开了另外一扇门，但是大家需要认真思考如何才能用好深层语义分析技术与深度学习技术。而随着社交媒体的广泛使用，我们也可充分利用社交媒体数据为我们的研究服务。

为了更好地推动文本到文本的生成技术的发展，业界可从以下几个方面着手：

其一，构建大规模评测数据集。数据是研究的基石，大规模、高质量的评测数据集对于研究工作至关重要，而目前上述多个任务均缺少大规模评测数据集，尤其是中文评测数据集。数据集的构建需要耗费大量人力物力，因此一个可行的途径就是采用众包的方式。

其二，构建开源平台。尽管针对上述各项任务业界均提出了多种解决方法，但很多方法并不易实现。业界需要为每个任务构建一个开源平台，将主流算法集成到该平台中，将会极大地方便后来者的研究，推动研究的发展。

3 意义到文本的生成

3.1 国际研究现状

不同于文本到文本的生成，意义到文本的生成这一任务的输入在学界并没有达成一致，其根本在于不论是哲学家还是语言学家，对何为自然语言的语义都未能形成较为一致的定义。在计算语言学领域，研究人员普遍遵循的语义研究原则建立在“真值条件（Truth Condition）”的基础上，认为寻找到了能够使自然语言语句成真的条件，即是在某种程度上刻画了自然语言的语义。在真值条件假设基础上，学者普遍采用逻辑的方法来对语义进行表征，并分别从模型论（Model Theory）和证明论（Proof Theory）两个角度来展开研究，很多学者也常常称这类型的语义为逻辑语义。目前已有的意义到文本的生成研究，普遍假设使用逻辑语义表征（以逻辑表达式为代表）作为输入，而以自然语言

语句作为输出，本文也围绕这些研究展开介绍。图 1 给出了一个基于类型 λ 演算进行语义表征的实例，在该例子中，问题的输入是一个 λ 表达式，而输出是一个英语句子。

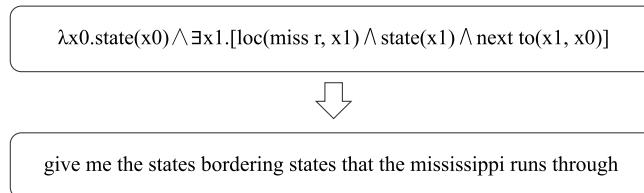


图 1 λ 表达式到文本的生成实例^[58]

意义到文本的生成和组合语义分析（Compositional Semantic Parsing）密切相关，语义分析旨在对线性的词序列进行自动句法语义解析并得到其真值条件。因为在分析过程中遵循了弗雷格所提之组合原则（Principle of Compositionality），因而称为组合语义分析，以与分布式语义（Distributional Semantics）相区别。组合语义分析是自然语言处理的一项核心技术，是迈向深度语义理解的一座重要桥梁，在多个自然语言处理核心任务中有着潜在应用，如智能问答、机器翻译等。从问题自身的定义来看，意义到文本的生成与组合语义分析是一对互逆的自然语言处理任务。在当前的国际研究中，仅专注于意义到文本的生成这一任务的学者并不多，部分以句法语义分析研究为主的学者会兼顾这方面的研究。

3.1.1 基于深层语法的文本生成

在早期的自然语言处理研究中，计算语言学发挥了很大的作用，计算语言学家从形式化、可计算的角度对自然语言进行建模，提出一系列的旨在解释语言运作机理的句法语义模型，并根据这些模型构建自然语言处理系统。相关研究在上个世纪八九十年代取得了丰硕的研究成果，一系列兼具语言本体解释力和可计算性的语法规范式（Grammar Formalism）被提出，如组合范畴语法（Combinatory Categorial Grammar, CCG）^[59] 和中心语驱动的短语结构语法（Head-driven Phrase-Structure Grammar, HPSG）^[60] 等。不同于目前句法分析所主要使用的上下文无关文法（Context-Free Grammar, CFG），上述语法规范式具有超越上下文无关的表达能力，其语法推导过程往往更复杂，蕴含更多的信息，而这些信息可以用来做更透明的语义分析。简单而言，这些深层语法规范式能够更好地支持句法语义同步的语言分析。在深层语法的支撑下，通过句法语义的协同推导可以获取自然语言的组合语义；而当以语义表征作为输入，通过其逆过程可以完成意义到文本的生成。

Shieber^[61]提出了一个统一的框架用于进行句法语义分析与生成。在这一框架中，Shieber 将语言处理统一理解为逻辑推演（Deduction）过程，其差别在于推演的始点（公理）与推演的终点（目标）不同。在这一视角下，传统的句法分析（Parsing）技术可以移植到文本生成上来，如线图分析法（Chart Parsing）技术可以转化为线图生成（Chart Generation）技术^[62]。Shieber 后续又同其他学者合作，将推演的思想细化，利用合一语法来表达句法语义接口（Syntax-Semantics Interface），提出了语义中心驱动的生成^[63]。

深层语法复杂度较高，如何构造对错综复杂的语言现象具有高覆盖度（Broad

Coverage) 的语法规则本身是一个极大的难题。以上研究主要是对原型算法进行讨论，而因为真实可用的大型深层语法当时没有得到很好的开发，以上研究并没有呈现极具代表性的经验结果。经过十余年的漫长开发，研究人员在 HPSG 理论的基础上开发了英语资源语法（English Resource Grammar, ERG [⊖]）^[64]，它是一个比较成功的具有较高覆盖率的深层语法规则系统，而围绕 ERG 所展开的文本生成研究也取得了有益的进展。Carroll 和 Oepen^[65] 基于 ERG 和真实测试数据重新讨论了基于线图的生成技术，给出了极具参考意义的经验评估；另外，他们也提出了两项新的技术来改进基于合一语法的可行解紧致表示（Compact Representation）及其相关解码算法——Selective Unpacking，尤其后者，有效地利用了判别式学习模型来改进文本生成过程中所遇到的歧义消解。

组合范畴语法是一个广受自然语言处理领域学者关注的语法范式，其设计遵循了类型透明（Type Transparency）的原则，具有精简的语法语义接口，常常被语义分析^[66]和文本生成^[67]模型所采用。White 和 Baldridge^[67] 讨论了如何将线图生成法与组合范畴语法结合，并开发了开源的基于组合范畴语法的句子实现（Realization）工具——OpenCCG [⊖]。White 又同其他学者联合提出了一些进一步改进文本生成的算法^[68-70]。

3.1.2 基于同步文法的文本生成

在过去的 20 年间，统计句法分析与统计机器翻译是公认的两个取得长足进步的自然语言处理任务。除了从成熟的统计句法分析中借鉴成功经验——如判别式消歧之外，不少学者也尝试复用成功的机器翻译模型来完成文本生成。机器翻译的目标是将某种自然语言语句翻译成另外一种自然语言语句，并尽量保持意义不变；而文本生成则可以视为将某种形式语言语句翻译成一种自然语言语句，二者具有极强的可比性。

Chiang^[71] 提出了层级基于短语的翻译模型（Hierarchical Phrase-based Model），其核心是利用同步上下文无关文法（Synchronous Context-Free Grammar）来协同源语言语句的解析和目标语言语句的生成。目前同步文法也已经被借鉴到文本生成的研究中^[72, 58]。Wong 与 Mooney^[72] 两位作者讨论了两种形式语言用于表征意义：第一种是用于指挥机器人动作的形式语言，第二种是一种无变量的数据库检索语言。而 Lu 与 Ng^[58] 则针对表达能力极强的类型 λ 表达式（Typed λ -expression）展开研究。两项研究的共同点是构建形式语言的基于树的结构，再将相关结构与待生成的自然语言的树结构建立一致性对应，从而完成文本生成任务。另一个共同点则是广泛地使用了现有的机器翻译技术（包括开源软件等）来进行文法抽取、解码等。

3.2 国内研究现状

国内语言学界与计算语言学界针对自然语言语义的形式化研究较少，针对汉语进行全面组合语义刻画的研究目前尚属空白。另一方面，从事自然语言处理的研究人员也较少涉猎深层语言结构处理问题，而对意义到文本的生成研究则更是鲜有，很少能见到

[⊖] [http://www.delph-in.net/erg/。](http://www.delph-in.net/erg/)

[⊖] [https://github.com/OpenCCG/openccg。](https://github.com/OpenCCG/openccg)

相关学术成果发表在重要学术会议和期刊上。

3.3 发展趋势与展望

随着深层自然语言理解的发展，研究者将越来越多的目光投向了意义到文本的生成这一自然语言生成核心任务上。意义到文本的生成这一任务，随着意义表征体系的不同，问题的复杂度也会随之变化，传统的基于深层语法分析的生成方法面临的解码效率差、语法鲁棒性不够等问题仍需要更好的技术解决方案。近些年来，有零星的一些工作尝试将较为成熟的组合优化技术应用到句法分析和机器翻译，如拉格朗日松弛^[73,74]，尝试去求解一些所涉及的 NP 难问题。应对意义到文本的生成这一复杂度高的问题，我们也可以尝试应用相关技术。而针对深层语法鲁棒性不够的问题，基于数据驱动的语法近似 (Grammar Approximation)^[75] 取得了不错的成绩，结果显示低阶语法近似能够有效改进深层语法分析的鲁棒性。如何应用相关思想来解决文本生成中所遇到的问题也是一个非常值得研究的方向。

而就针对汉语的文本生成研究来说，需要国内外学界做出更大的努力。首先，在语言本体分析方面，需要学者们建立相关的语义表征体系及针对汉语的特殊语言现象的分析，以支持汉语的深层处理。其次，在文本生成算法方面，也需要我们投入更多的科研精力，设计适合汉语自动生成的模型算法等。

4 数据到文本的生成

4.1 国际研究现状

数据到文本的生成技术指根据给定的数值数据生成相关文本，例如基于数值数据生成天气预报文本、体育新闻、财经报道、医疗报告等。数据到文本的生成技术具有极强的应用前景，目前该领域已经取得了很大的研究进展，业界已经研制出面向不同领域和应用的多个生成系统。针对数据到文本的生成技术的研究单位主要集中在少数几个单位，例如英国阿伯丁大学、布莱顿大学、爱丁堡大学等，相关研究成果主要发表在 INLG、ENLG 这几个专业学术会议上。

英国阿伯丁大学的 Ehud Reiter 在三阶段流水线模型^[76]的基础上提出了数据到文本的生成系统的一般框架，如图 2 所示。

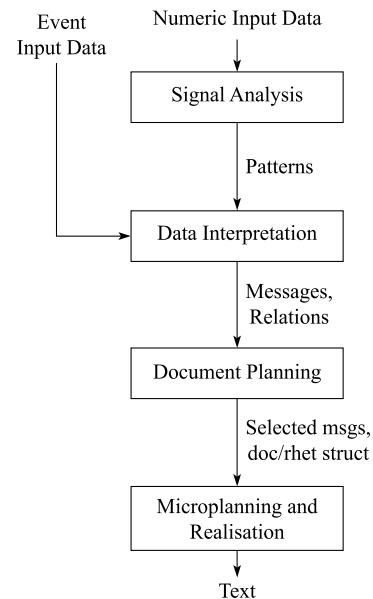


图 2 数据到文本的生成系统的一般框架^[77]

其中：

信号分析（Signal Analysis）模块的输入为数值数据，通过利用各种数据分析方法检测数据中的基本模式，输出离散数据模式。例如股票数据中的峰值、较长期的增长趋势等。该模块与具体应用领域和数据类型相关，不同的应用领域与数据类型所输出的数据模式是不同的。

数据阐释（Data Interpretation）模块的输入为基本模式与事件，通过对基本模式和输入事件进行分析，推断出更加复杂和抽象的消息，同时推断出它们之间的关系，最后输出高层消息以及消息之间的关系。例如针对股票数据，如果跌幅超过某个值则可以创建一条消息。还需要检测消息之间的关系，例如因果关系、时序关系等。值得说明的是，数据阐释模块并不是在所有文本生成系统中都需要，例如，在天气预报文本生成系统中，基本的模式足以满足要求，因此并不需要采用数据阐释模块。

文档规划（Document Planning）模块的输入为消息及关系，分析并决定哪些消息和关系需要在文本中提及，同时要确定文档结构，最后输出需要提及的消息以及文档结构。从更高的层次来说，信号分析与数据阐释模块会产生大量的消息、模式和事件，但文本通常长度受限，只能描述其中的一部分，因此文档规划模块必须确定文本中需要说明的消息。一般可根据专家知识以及消息的重要性、新颖性等来进行选择和确定。当然，该模块也与领域有很大关系，不同领域，消息选择时所考虑的因素不一样，文档的结构也会不一样。

微规划与实现（Microplanning and Realisation）模块的输入为选中的消息及结构，通过自然语言生成技术输出最终的文本。该模块主要涉及对句子进行规划以及句子实现，要求最终实现的句子具有正确的语法、形态和拼写，同时采用准确的指代表达。所采用的技术在学术界有相当多的研究，具体可参考本文第3节“意义到文本的生成”。

目前，业界已经研制了面向多个领域的数据到文本的生成系统，这些系统的框架与上述一般框架并无大的差别，部分系统将上述框架中的两个模块合并为一个模块，或者省去了其中一个模块。

数据到文本的生成技术在天气预报领域应用最为成功，业界研制了多个系统对天气预报数据进行总结，生成天气预报文本。例如，FoG 系统^[78]能够从用户操作过的数据中生成双语天气预报文本；SumTime 系统^[79]能够生成海洋天气预报文本，实验评测表明用户有时候更倾向于阅读 SumTime 所生成的天气预报，而非专家撰写的天气预报^[80]。此外，英国阿伯丁大学的 AnjaBelz 提出概率生成模型进行天气语言文本的生成^[81]。AnjaBelz 和 Eric Kow 进一步基于天气预报数据分析对比了多种数据到文本的生成系统，结果表明采用自动化程度较高的方法并不会降低文本生成质量，同时文本质量的自动评价方法会低估基于手工规则构建的系统，而高估自动化系统^[82]。

业界面向其他领域也研制了多个文本生成系统，例如针对空气质量的文本生成系统^[83]，针对财经数据的文本生成系统^[84]，面向医疗诊断数据的文本生成系统 TOPAZ^[85]、Suregen^[86]、BT-45^[87]等。其中 BT-45 能够为新生儿重症监护病房（NICU）的监控数据生成文本摘要，帮助医生进行决策。图 3 和图 4 分别给出了 BT-45 系统的输入样例与生成得到的文本。

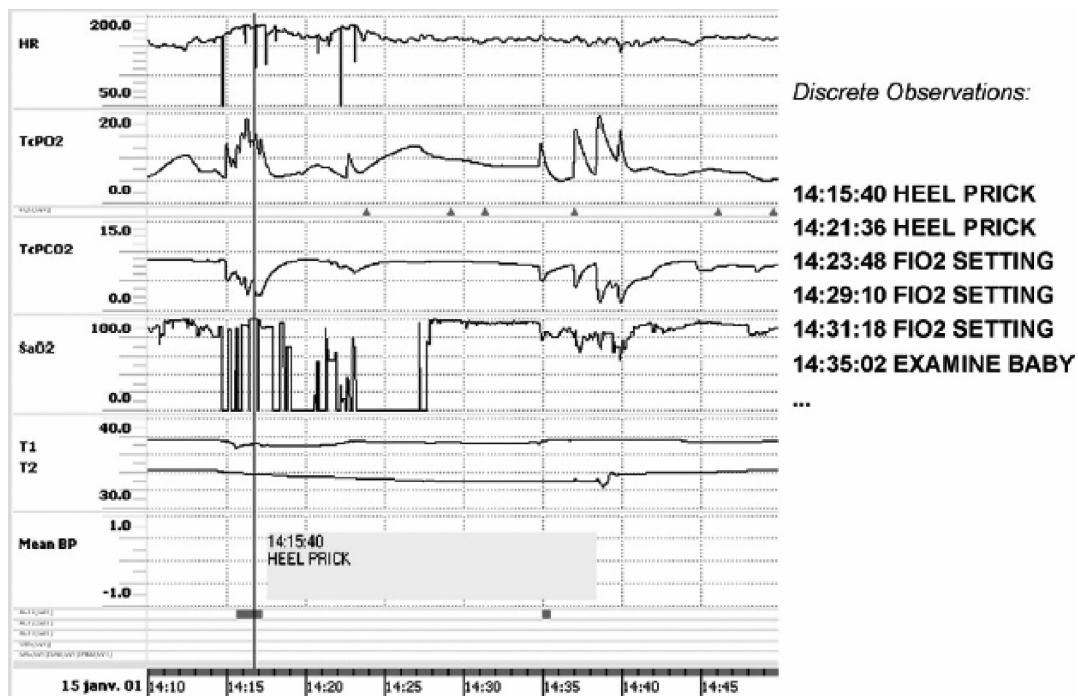


图3 NICU 数据样例，从上到下分别表示 HR、TcPO2、TcPCO2、SaO2、T1 & T2 和 Mean BP^[87]

You saw the baby between 14:10 and 14:50. Heart Rate (HR) = 159. Core Temperature (T1) = 37.7. Peripheral Temperature (T2) = 34.3. Transcutaneous Oxygen (TcPO2) = 5.8. Transcutaneous CO2 (TcPCO2) = 8.5. Oxygen Saturation (SaO2) = 89.
Over the next 30 minutes T1 gradually increased to 37.3.
By 14:27 there had been 2 successive desaturations down to 56. As a result, Fraction of Inspired Oxygen (FIO2) was set to 45%. Over the next 20 minutes T2 decreased to 32.9. A heel prick was taken. Previously the spo2 sensor had been re-sited.
At 14:31 FIO2 was lowered to 25%. Previously TcPO2 had decreased to 8.4. Over the next 20 minutes HR decreased to 153.
By 14:40 there had been 2 successive desaturations down to 68. Previously FIO2 had been raised to 32%. TcPO2 decreased to 5.0. T2 had suddenly increased to 33.9. Previously the spo2 sensor had been re-sited. The temperature sensor was re-sited.

图4 BT-45 系统生成的对应文本^[87]

由于数据到文本的生成技术的巨大应用价值，工业界成立了多家从事文本生成的公司，能够为多个行业基于行业数据生成行业报告或新闻报道，从而节省大量的人力。比较知名的公司有 ARRIA^①、AI^②、NarrativeScience^③等。其中 ARRIA 是一家总部设在欧洲的公司，其前称为 Data2Text，由来自阿伯丁大学的两名教授 Ehud Reiter 与 YajiSripada 创办，后来自然语言生成领域的另一位科学家 Robert Dale 也加入了该公司，该公司的核心技术为 ARRIA NLG 引擎。AI (Automated Insights) 则是一家美国人工智能公司，由一名思科的前工程师 Robbie Allen 所创办，最早基于体育数据生成文本摘要，目前能为包括金融、个人健身、商业智能、网站分析等在内的多个领域内的数据生成文本报告，其核心技术为 WordSmith NLG 引擎。目前，AI 公司已经为美联社等多家单位生成数亿篇新闻

① <https://www.arria.com/>。

② <http://automatedinsights.com>。

③ <http://www.narrativescience.com>。

报道，产生了巨大的影响力。NarrativeScience 则是根据美国西北大学的一个研究项目 StatsMonkey 发展而来，其核心技术为 Quill NLG 引擎。Forbes 是 NarrativeScience 的一个典型客户，在网站上有个 NarrativeScience 专页[⊖]，其中全部文章都是由 NarrativeScience 自动生成。图 5 给出一篇自动生成的样例新闻。

The screenshot shows a news article from Forbes. At the top left, it says 'INVESTING' and the date '7/07/2015 @ 1:00下午 | 332 views'. The main title is 'Earnings for Alcoa Projected to Rise'. Below the title, it says 'By Narrative Science'. There are two small links: '+ Comment Now' and '+ Follow Comments'. The first paragraph discusses Wall Street's expectations for Alcoa's earnings. The second paragraph discusses the company's revenue projection for the fiscal year. The third paragraph provides background on Alcoa's industry and other companies in the sector. The final sentence at the bottom is a note from Zacks.

INVESTING 7/07/2015 @ 1:00下午 | 332 views

Earnings for Alcoa Projected to Rise

By Narrative Science

+ Comment Now + Follow Comments

Wall Street is high on **Alcoa**, expecting it to report earnings that are up 28% from a year ago when it reports its second-quarter earnings on Wednesday, July 8, 2015. The consensus estimate is 23 cents per share, up from earnings of 18 cents per share a year ago.

The consensus estimate has fallen over the past three months, from 27 cents. Analysts are expecting earnings of 95 cents per share for the fiscal year. Analysts look for revenue to decrease 1% year-over-year to \$5.79 billion for the quarter, after being \$5.84 billion a year ago. For the year, revenue is projected to roll in at \$23.63 billion.

Revenue dropped year-over-year in the first quarter, ending a two-quarter streak of growing revenue.

Alcoa is a global producer of aluminum. It is mainly engaged in the production and management of primary aluminum, fabricated aluminum, and alumina combined. It is actively involved in a range of industries, including technology, mining, smelting, and recycling. Kaiser Aluminum Corp., also in the metal mining industry, will report earnings on Wednesday, July 22, 2015. Analysts are expecting earnings of \$1.19 per share for Kaiser Aluminum, up 13% from last year's earnings of \$1.05 per share. Other companies in the metal mining industry with upcoming earnings release dates include: Noranda Aluminum Holding and Aluminum Corp. of China Limited.

Earnings estimates provided by Zacks.

图 5 NarrativeScience 自动生成的样例新闻

4.2 国内研究现状

国内学术界对数据到文本的生成鲜有研究，也很少见到相关学术成果发表在重要学术会议和期刊上。国内工业界则有部分单位研制了基于模板的文本生成系统。例如新华

⊖ <http://www.forbes.com/sites/narrativescience>。

社已开发了从财报数据生成企业财报年报的系统，该系统基于人工模板，将需要的数据填入写好的模板中，从而生成财报年报。由于采用的模板比较固定，所以为不同企业生成的财报年报都比较类似，不够生动。

4.3 发展趋势与展望

从数据到中文文本的生成技术很有研究意义，同时实用性很强。如果能实现从数据到中文新闻的生成，那么将极大缓解编辑和记者的负担，实现媒体、出版行业的变革。而实现这样的系统，必须依靠科研院所和新闻出版机构的合作，新闻出版机构能够提供大量的数据和专家知识，而科研院所则擅长自然语言理解与生成的理论与方法。

此外，要研制一套通用的面向不同领域的数据到文本的生成系统相当复杂和困难，因此一个更好的做法是先选择一两个领域（如财经、体育）进行系统研制，待系统成熟后再考虑将系统迁移到其他领域。

5 图像到文本的生成

5.1 国际研究现状

图像到文本的生成技术是指根据给定的图像生成描述该图像内容的自然语言文本，例如新闻图像附带的标题、医学图像附属的说明、儿童教育中常见的看图说话，以及用户在微博等互联网应用中上传图片时提供的说明文字。依据所生成自然语言文本的详细程度及长度的不同，这项任务又可以分为图像标题自动生成和图像说明自动生成。前者需要根据应用场景突出图像的核心内容，例如，为新闻图片生成的标题需要突出与图像内容密切关联的新闻事件，并在表达方式上求新以吸引读者的眼球；而后者通常需要详细描述图像的主要内容，例如，为有视力障碍的人提供简洁翔实的图片说明，力求将图片的内容全面且有条理地陈述出来，而在具体表达方式上并没有具体的要求。

对于图像到文本的自动生成这一任务，人类可以毫不费力地理解图像内容，并按具体需求以自然语言句子的形式表述出来；然而对于计算机而言，则需要综合运用图像处理、计算机视觉和自然语言处理等几大领域的研究成果。作为一项标志性的交叉领域研究任务，图像到文本的自动生成吸引着来自不同领域研究者的关注。自 2010 年起，自然语言处理界的知名国际会议和期刊 ACL、TACL 和 EMNLP 中都有相关论文的发表，而自 2013 年起，模式识别与人工智能领域顶级国际期刊 IEEE TPAMI 以及计算机视觉领域顶级国际期刊 IJCV 也开始刊登相关工作的研究进展，至 2015 年，计算机视觉领域的知名国际会议 CVPR 中更是有近 10 篇相关工作的论文发表，同时机器学习领域知名国际会议

ICML 中也有两篇相关论文发表。图像到文本的自动生成任务已被认为是人工智能领域中的一项基本挑战。

与一般的文本生成问题类似，解决图像到文本的自动生成问题也需要遵循三阶段流水线模型^[76]，同时又需要根据图像内容理解的特点做出一些调整：

在内容抽取方面，需要从图像中抽取物体、方位、动作、场景等概念，其中物体可以具体定位到图像中的某一具体区域，而其他概念则需要进行语义标引。这部分主要依靠模式识别和计算机视觉技术。

在句子内容选择方面，需要依据应用场景选择最重要（如图像画面中最突出的，或与应用场景最相关的）且意义表述连贯的概念。这部分需要综合运用计算机视觉与自然语言处理技术。

最后，在句子实现部分，根据实际应用特点选取适当的表述方式将所选择的概念梳理为合乎语法规则的自然语言句子。这部分主要依靠自然语言处理技术。

早期工作主要依照上述三阶段的流水线模式来实现。例如，在 Yao 等人^[88]的工作中，图像被细致地分割并标注为物体及其组成部分，以及图像所表现的场景，并在此基础上选择与场景相关的描述模板，将物体识别的结果填充到模板得到图像的描述文字。而 Feng 与 Lapata^[89,90]则采用概率图模型对文本信息和图像信息同时建模，并从新闻图片所针对的文字报道中挑选合适的关键词作为体现图像内容的关键词，并进而利用语言模型将所选取的内容关键词及必要的功能词汇链接为基本合乎语法规则的图像标题。还有一些工作^[91-95]则依靠计算机视觉领域现有的物体识别技术从图像中抽取物体（包括人物、动物、花草、车、桌子等常见的物体类型），并对其定位以获得物体之间的上下位关系，进而依赖概率图模型和语言模型选取适当的描述顺序，将这些物体概念、介词短语块串联成完整的句子。Hodosh 等人^[96]则利用基于核函数的典型关联分析（Kernel Canonical Correlation Analysis, KCCA）来寻找文本与图像之间的关联，并依据图像信息对候选句子排序，从而获得最佳描述句子。值得说明的是，Hodosh 等人^[96]的工作和 Feng 与 Lapata^[89,91]的工作均没有依靠现有的物体识别技术。

图 6 给出了一种典型的流水线模型。

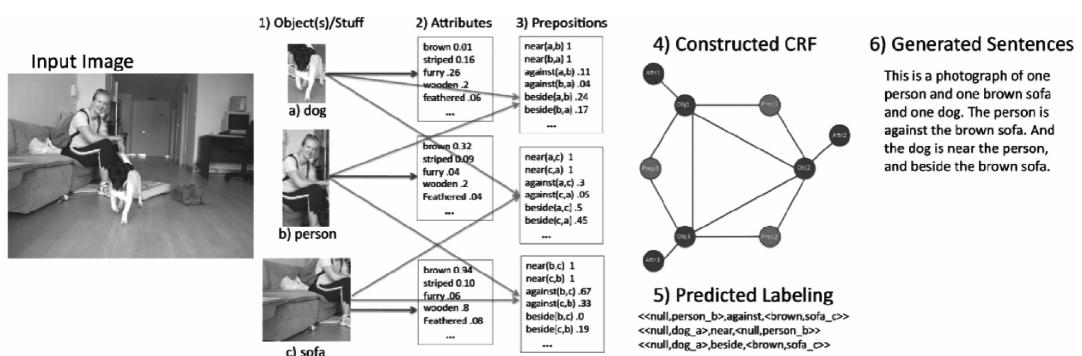


图 6 一种典型的流水线模型^[93]

随着深度学习方法在模式识别、计算机视觉及自然语言处理领域的广泛应用，基于海量数据的大规模图像分类、语义标注技术得到了快速发展；同时，统计机器翻译等与自然语言生成相关的技术也有了显著的提高。这也催生了将图像语义标注及自然语言句子生成进行联合建模的一系列工作，一方面在图像端采用多层深度卷积神经网络（Deep Convolution Neural Network, DCNN）对图像中的物体概念进行建模，另一方面在文本端采用循环神经网络（Recurrent Neural Network, RNN）或递归神经网络（Recursive Neural Network）对自然语言句子的生成过程进行建模^[97]。传统图像语义标注工作主要关注某个具体物体的识别以及物体之间的相对位置关系，而对动作等抽象概念的关注较少。Socher 等人^[98]提出利用递归神经网络对句子建模，并利用句法解析树突出对于动作（动词）的建模，进而将图像端与文本端进行联合优化，较好地刻画了物体与动作之间的关系。为了将两种不同模态的数据统一在一个框架下，Chen 与 Zitnick^[99]将文本信息与图像信息融合在同一个循环神经网络中，利用图像信息作为记忆模块，从而指导文本句子的生成，同时又借助于一个重构图像信息层，实现了图像到文本、文本到图像的双方向表示。而 Mao 等人^[100]则通过 DCNN 得到的图像信息与文本信息融合到同一个循环神经网络（m-RNN）中，将图像信息融入到自然语言句子生成的序列过程中，取得了不错的结果。类似的想法也被 Donahue 等人^[101]应用于动作识别和视频描述生成过程中。但在 m-RNN 的句子生成过程中，在图像端并没有显著的约束，例如在图 7 中，当生成单词“man”的时候，并没有与图像信息中的任务标注发生直接或间接的关联。

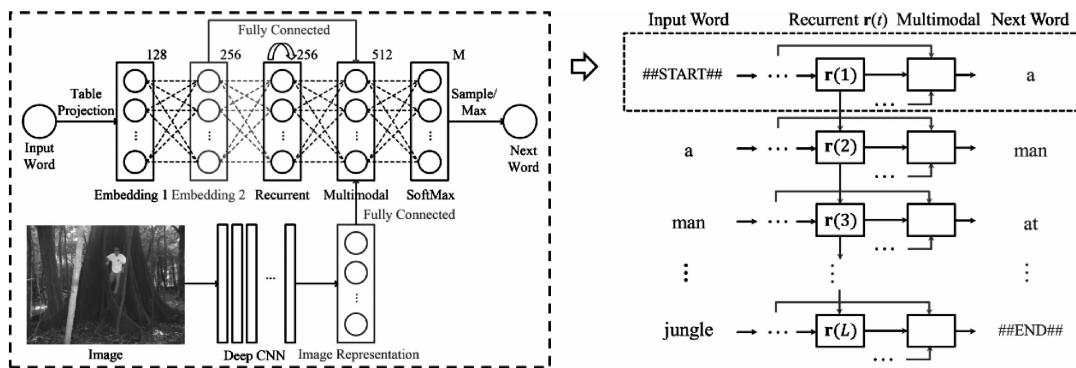


图 7 多模态 m-RNN 模型^[100]

谷歌与加拿大蒙特利尔大学和多伦多大学的研究人员则分别借鉴了统计机器翻译领域的最新研究进展来推进图像到文本自动生成的联合建模^[102,103]。前者利用深层卷积神经网络对图像建模，将图像信息“编码”后，直接由另一个与之相连接的 LSTM（Long-Short Term Memory Network）神经网络“解码”成自然语言句子，无须进行图像-词对齐、调序等传统模型的子步骤。而后者则在基于神经网络的机器翻译框架下，提出利用计算机视觉领域中的“注意”（Attention）机制来促进词语和图像块之间的对齐，从而在句子生成过程中，模拟人视觉的“注意”转移过程能够与词语序列的生成过程相互促进，使生成的句子更符合人的表述习惯。图 8 给出一个视觉“注意”引导的图像标题生

成过程例子。

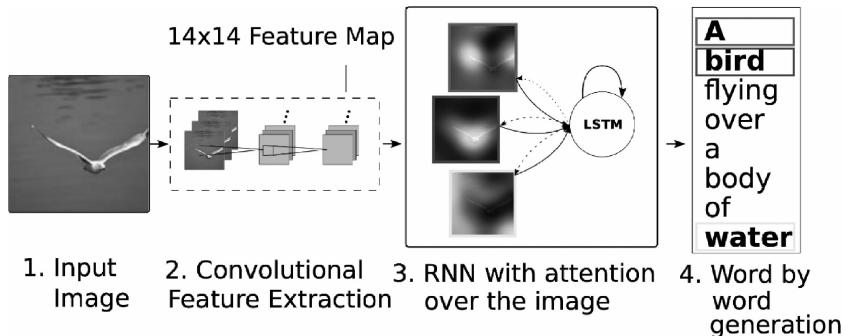


图 8 视觉“注意”引导的图像标题生成过程^[103]

此外，微软的研究人员^[104]利用卷积神经网络 CNN 和多示例学习（Multiple Instance Learning, MIL）对图像建模，并利用判别式语言模型生成候选句子，并采用统计机器翻译研究中经典的最小误差率训练（Minimum Error Rate Training, MERT）来发掘文本和图像层面的特征对候选句子进行排序。

虽然图像到文本的生成技术还处在探索阶段，距离实际产业应用还有一定的距离，但工业界已经开始注意到这一技术的理论研究价值和潜在应用前景，积极与学术界合作拓展研究方向。在 2015 年的计算机视觉知名国际会议 CVPR 2015 上举办的 LSUN（Large-scale Scene UNderstanding）挑战活动中也进行了图像标题自动生成的评测任务，最终谷歌公司^[102]和微软研究院^[104]取得了总成绩并列第一名，蒙特利尔 - 多伦多联队^[103]和另一只微软研究院队伍^[105]总成绩并列第三名，加州伯克利分校^[101]获得第五名。

5.2 国内研究现状

国内学术界对图像到文本的生成技术研究开展较晚，大部分科研单位专注于跨媒体数据的语义标注和检索等任务上，只有人大、清华、北大、北航和中科院等科研单位开展了相关研究，如人民大学与腾讯合作在 2015 年欧盟组织的 ImageCLEF 评测中，在图像句子生成（Image Sentence Generation）任务中取得了第一名。

在工业界方面，百度和腾讯等科研机构也依靠自身在跨媒体语义标注、分类和检索等方面的研究优势，逐步开展相关方向的研究工作，如百度与 UCLA 合作的 m-RNN 系统在 CVPR 2015 LSUN 评测的图像标题自动生成任务中也取得了不错的成绩。

5.3 发展趋势与展望

从图像到文本的生成技术需要集成模式识别与机器学习、计算机视觉、自然语言处

理，甚至认知科学领域的研究成果，具有极高的理论研究价值和实用前景。从一定程度上讲，这一技术同图像语义标注等任务一道，已成为各大顶尖科研机构在人工智能领域综合研究实力的较量方式，必将促进其快速发展。

而对于这一任务本身而言，更大的挑战仍在于如何正确地抽取图像的内容，同时根据人类的语言习惯选择适当的表述方式将图像内容转换为自然语言句子。需要指出的是，目前的研究仍然聚焦在是否将图像中的物体概念抽取完全，是否选择了正确的词语，所生成的句子是否符合语法习惯等。可以预见在不久的将来，实际应用场景和上下文语境等约束将进一步推进相关技术的进步，必将广泛应用于新闻传播、在线教育、智能家居等多个领域。

6 总结与展望

本文对文本自动生成技术进行了全面的介绍，包括文本到文本的生成、意义到文本的生成、数据到文本的生成、图像到文本的生成等。由于上述每项技术均有众多的研究者在研究，相关的学术成果也层出不穷，因此本文的总结难免有遗漏之处。希望本文的内容能够对相关研究人员和从业者有所帮助。

对文本自动生成技术的国际研究现状和国内研究现状进行比较可以看到，国内对该领域的研究投入和产出均远远不够，原创性的方法、资源、系统以及产品都相对比较匮乏，而且该领域没有受到业界足够的关注。我们必须奋起直追，建设相关中文资源，提出原创性文本生成方法，搭建中文文本生成系统并开发相关产品，才能占领中文文本生成的制高点。我们期待第一个中文文本生成系统由国内单位研制而成。

注：本文第1、2、4、6节由万小军撰写，第3节由孙薇薇撰写，第5节由冯岩松撰写。博士生姚金戈参与了校对工作。

参考文献

- [1] LuhnH P. The automatic creation of literature abstracts [J]. IBM Journal of research and development, 1958, 2(2) : 159-165.
- [2] LinC Y, & Hovy E. From single to multi-document summarization: A prototype system and its evaluation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics [C]. Association for Computational Linguistics, 2002 : 457-464.
- [3] Evans D K, Klavans J L, & McKeown K R. Columbia newsblaster: multilingual news summarization on the Web. Demonstration Papers at HLT- NAACL 2004 [C]. Association for Computational Linguistics, 2004 : 1-4.

- [4] Radev D, Otterbacher J, Winkel A, & Blair-Goldensohn, S. NewsInEssence: summarizing online news topics[J]. Communications of the ACM, 2005 , 48(10) , 95-98.
- [5] ShenD, Sun J T, Li H, Yang Q, & Chen, Z. Document Summarization Using Conditional Random Fields. IJCAI[C]. 2007 , 7: 2862-2867.
- [6] Conroy JM, & O'leary D P. Text summarization via hidden markov models. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval[C]. ACM, 2001 : 406-407.
- [7] Schilder F, & Kondadadi R. FastSum: fast and accurate query-based multi-document summarization. Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers[C]. Association for Computational Linguistics, 2008 : 205-208.
- [8] Ouyang Y, Li W, Li S, & Lu Q. Applying regression models to query-focused multi-document summarization[J]. Information Processing & Management, 2011 , 47(2) : 227-237.
- [9] Cao Z, Wei F, Dong L, Li S, & Zhou M. Ranking with recursive neural networks and its application to multi-document summarization. Twenty-Ninth AAAI Conference on Artificial Intelligence[C]. 2015.
- [10] Carbonell J, & Goldstein J. The use of MMR, diversity-based reranking for reordering documents and producing summaries. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval[C]. ACM, 1998 : 335-336.
- [11] Bollegala D, Okazaki N, & Ishizuka M. A bottom-up approach to sentence ordering for multi-document summarization[J]. Information processing & management, 2010 , 46(1) : 89-109.
- [12] McDonald R. A study of global inference algorithms in multi-document summarization [M]. Berlin Heidelberg: Springer, 2007 : 557-564.
- [13] Gillick D, & Favre B. A scalable global model for summarization. Proceedings of the Workshop on Integer Linear Programming for Natural Language Processing [C]. Association for Computational Linguistics, 2009 : 10-18.
- [14] LiC, Qian X, & Liu Y. Using Supervised Bigram-based ILP for Extractive Summarization[J]. ACL, 2013 , (1) : 1004-1013.
- [15] LinH, & Bilmes J. Multi-document summarization via budgeted maximization of submodular functions. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics[C]. Association for Computational Linguistics, 2010 : 912-920.
- [16] LinH, & Bilmes J. A class of submodular functions for document summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies[C]. Association for Computational Linguistics, 2011 , 1: 510-520.
- [17] Qian X, & Liu Y. Fast Joint Compression and Summarization via Graph Cuts. EMNLP [C]. 2013 : 1492-1502.
- [18] Li C, Liu Y, Liu F, Zhao L, & Weng F. Improving Multi-documents Summarization by Sentence Compression based on Expanded Constituent Parse Trees. EMNLP[C]. 2014.
- [19] Berg-Kirkpatrick T, Gillick D, & Klein D. Jointly learning to extract and compress. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies [C]. Association for Computational Linguistics, 2011 , 1: 481-490.
- [20] BarzilayR, & McKeown K R. Sentence fusion for multidocument news summarization[J]. Computational Linguistics, 2005 , 31(3) : 297-328.
- [21] Bing L, Li P, Liao Y, Lam W, Guo W, & Pasconneau R J. Abstractive Multi-Document Summarization

- via Phrase Selection and Merging. ACL[C]. 2015.
- [22] LiuF, Flamigan J, Thomson S, Sadeh N, & Smith N A. Toward Abstractive Summarization Using Semantic Representations. NAACL[C]. 2015.
- [23] Abu-Jbara A, & Radev D. Coherent citation-based summarization of scientific papers. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies [C]. Association for Computational Linguistics, 2011, 1: 500-509.
- [24] SaifMohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishnan, Vahed Qazvinian, Dragomir Radev, and David Zajic. Using citations to generate surveys of scientific paradigms. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics[C]. Association for Computational Linguistics, 2009: 584-592.
- [25] Nenkova A, & McKeown K. A survey of text summarization techniques[J]. Mining Text Data . Springer US, 2012: 43-76.
- [26] Knight K, & Marcu D. Summarization beyond sentence extraction: A probabilistic approach to sentence compression[J]. Artificial Intelligence, 2002, 139(1) : 91-107.
- [27] CohnT, & Lapata M. Sentence compression beyond word deletion. Proceedings of the 22nd International Conference on Computational Linguistics [C]. Association for Computational Linguistics, 2008, 1: 137-144.
- [28] Knight K, & Marcu D. Statistics-based summarization-step one: Sentence compression. AAAI/IAAI[C]. 2000 : 703-710.
- [29] McDonald R T. Discriminative Sentence Compression with Soft Syntactic Evidence. EACL[C]. 2006.
- [30] Cohn T A, & Lapata M. Sentence compression as tree transduction[J]. Journal of Artificial Intelligence Research, 2009: 637-674.
- [31] Clarke J, & Lapata M. Global inference for sentence compression: An integer linear programming approach [J]. Journal of Artificial Intelligence Research, 2008: 399-429.
- [32] Thadani K, & McKeown K. Supervised sentence fusion with single-stage inference. Proceedings of the Sixth International Joint Conference on Natural Language Processing[C]. 2013: 1410-1418.
- [33] Elsner M, & Santhanam D. Learning to fuse disparate sentences. Proceedings of the Workshop on Monolingual Text-To-Text Generation[C]. Association for Computational Linguistics, 2011: 54-63.
- [34] Filippova K. Multi-sentence compression: finding shortest paths in word graphs. Proceedings of the 23rd International Conference on Computational Linguistics [C]. Association for Computational Linguistics, 2010: 322-330.
- [35] Barzilay R, & Lee L. Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology[C]. Association for Computational Linguistics, 2003: 16-23.
- [36] Fujita A, Inui K, & Matsumoto Y. Exploiting lexical conceptual structure for paraphrase generation. IJCNLP 2005[C]. 2005, LNAI 3651: 908-919.
- [37] Quirk C, Brockett C, & Dolan W B. Monolingual Machine Translation for Paraphrase Generation. EMNLP [C]. 2004: 142-149.
- [38] Duboue P A, & Chu-Carroll J. Answering the question you wish they had asked: The impact of paraphrasing for question answering. Proceedings of the Human Language Technology Conference of the

- NAACL, Companion Volume: Short Papers[C]. Association for Computational Linguistics, 2006 : 33-36.
- [39] MaxA. Sub-sentential paraphrasing by contextual pivot translation. Proceedings of the 2009 Workshop on Applied Textual Inference[C]. Association for Computational Linguistics, 2009 : 18-26.
- [40] WubbenS, Van Den Bosch A, & Krahmer E. Sentence simplification by monolingual machine translation. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers [C]. Association for Computational Linguistics, 2012 , 1: 1015-1024.
- [41] ZhuZ, Bernhard D, & Gurevych I. A monolingual tree-based translation model for sentence simplification. Proceedings of the 23rd international conference on computational linguistics [C]. Association for Computational Linguistics, 2010: 1353-1361.
- [42] Woodsend K, & Lapata M. WikiSimple: Automatic Simplification of Wikipedia Articles. AAAI[C]. 2011.
- [43] Wan X, Yang J, & Xiao J. Manifold-Ranking Based Topic-Focused Multi-Document Summarization. IJCAI [C]. 2007 , 7: 2903-2908.
- [44] WanX, & Yang J. Multi-document summarization using cluster-based link analysis. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval[C]. ACM, 2008 : 299-306.
- [45] WanX, & Zhang J. CTSUM: extracting more certain summaries for news articles. Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval [C]. ACM, 2014 : 787-796.
- [46] YanS, & Wan X. SRRank: leveraging semantic roles for extractive multi-document summarization[J]. Audio, Speech, and Language Processing, IEEE/ACM Transactions on, 2014, 22(12) : 2048-2058.
- [47] Jin-ge Yao, Xiaojun Wan, Jianguo Xiao. Compressive Document Summarization via Sparse Optimization. IJCAI[C]. 2015.
- [48] YanR, Wan X, Otterbacher J, Kong L, Li X, & Zhang Y. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval[C]. ACM, 2011 : 745-754.
- [49] WanX. Using bilingual information for cross-language document summarization. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies[C]. Association for Computational Linguistics, 2011 , 1: 1546-1555.
- [50] Wan X, Jia H, Huang S, & Xiao J. Summarizing the differences in multilingual news. Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval[C]. ACM, 2011 : 735-744.
- [51] HuY, & Wan X. PPSGen: Learning- Based Presentation Slides Generation for Academic Papers [J]. Knowledge and Data Engineering, IEEE Transactions on, 2015 , 27(4) : 1085-1097.
- [52] HuY, & Wan X. Automatic Generation of Related Work Sections in Scientific Papers: An Optimization Approach. EMNLP[C]. 2014.
- [53] YaoJ G, Wan X, & Xiao J. Joint Decoding of Tree Transduction Models for Sentence Compression. EMNLP[C]. 2014.
- [54] HuangM, Shi X, Jin F, & Zhu X. Using first-order logic to compress sentences. Twenty-Sixth AAAI Conference on Artificial Intelligence[C]. 2012.
- [55] ShiqiZhao, Cheng Niu, Ming Zhou, Ting Liu, and Sheng Li. Combining Multiple Resources to Improve SMT- based Paraphrasing Model. Proceedings of the 46th Annual Meeting of the Association for

- Computational Linguistics: Human Language Technologies (ACL-08: HLT) [C]. 2008: 1021-1029.
- [56] Shiqi Zhao, Haifeng Wang, Xiang Lan, and Ting Liu. Leveraging Multiple MT Engines for Paraphrase Generation. Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010) [C]. 2010: 1326-1334.
- [57] Shiqi Zhao, Xiang Lan, Ting Liu, Sheng Li. Application- driven Statistical Paraphrase Generation. Proceedings of Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009) [C]. 2009: 834-842.
- [58] WeiLu, Hwee Tou Ng. A Probabilistic Forest-to-String Model for Language Generation from Typed Lambda Calculus Expressions. Proceedingds of the 2011 Conference on Empirical Methods in Natural Language Processing[C]. 2011.
- [59] MarkSteedman. The Syntactic Process[M]. MIT Press, 2000.
- [60] Carl Pollard, Ivan A Sag. Head- Driven Phrase Structure Grammar [M]. University of Chicago Press, 1994.
- [61] StuartMShieber. A uniform architecture for parsing and generation. Proceedings of the 12th International Conference on Computational Linguistics[C]. 1988.
- [62] Martin Kay. Chart Generation. Proceedings of the 34th annual meeting on Association for Computational Linguistics[C]. 1996.
- [63] Stuart M Shieber, Gertjan van Noord, Fernando C N Pereira, and Robert C Moore. Semantic-head-driven generation[J]. Computational Linguistics, 1990.
- [64] DanFlickinger. On building a more efficient grammar by exploiting types [J]. Collaborative Language Engineering, 2002.
- [65] CarrollJ, & Oopen S. High efficiency realization for a wide- coverage unification grammar. Natural Language Processing-IJCNLP 2005[C]. Berlin Heidelberg: Springer, 2005: 165-176.
- [66] Luke S Zettlemoyer and Michael Collins. Learning to Map Sentences to Logical Form: Structured Classification with Probabilistic Categorial Grammars. Proceedings of UAI[C]. 2005.
- [67] Michael White and Jason Baldridge. Adapting Chart Realization to CCG. Proceedings of the 9th European Workshop on Natural Language Generation[C]. 2003.
- [68] Michael White. Reining in CCG Chart Realization. Proceedings of the 3rd International Conference on Natural Language Generation[C]. 2004.
- [69] Michael White. CCG Chart Realization from Disjunctive Inputs. Proceedings of the 4th International Conference on Natural Language Generation (INLG-06) [C]. 2006.
- [70] Michael White, Rajakrishnan Rajkumar, and Scott Martin. Towards Broad Coverage Surface Realization with CCG. Proceedings of the 2007 Workshop on Using Corpora for NLG: Language Generation and Machine Translation[C]. 2007.
- [71] David Chiang. A Hierarchical Phrase- Based Model for Statistical Machine Translation. Proceedings of the 43rd annual meeting on Association for Computational Linguistics[C]. 2005.
- [72] YukWah Wong, Raymond Mooney. Generation by Inverting a Semantic Parser that Uses Statistical Machine Translation. Proceedings of Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics[C]. 2007.
- [73] Terry Koo, Alexander M Rush, Michael Collins, Tommi Jaakkola, and David Sontag. Dual Decomposition

- for Parsing with Non-Projective Head Automata. Proceedings of EMNLP 2010[C]. 2010.
- [74] Alexander M Rush and Michael Collins. Exact Decoding of Syntactic Translation Models through Lagrangian Relaxation. Proceedings of ACL 2011[C]. 2011.
- [75] YiZhang, Hans- Ulrich Krieger. 2011. Large- Scale Corpus- Driven PCFG Approximation of an HPSG. In*Proceedings of 12th International Conference on Parsing Technologies*.
- [76] Reiter E & Dale R. Building natural language generation systems (Vol. 33) [M]. Cambridge : Cambridge university press.
- [77] Reiter E. An architecture for data- to- text systems. Proceedings of the Eleventh European Workshop on Natural Language Generation[C]. Association for Computational Linguistics , 2007 : 97-104.
- [78] Goldberg E, Driedger N, & Kittredge R. Using natural- language processing to produce weather forecasts [J]. IEEE Expert , 1994 , 9(2) : 45-53.
- [79] Sripada S, Reiter E, & Davy I. SumTime- Mousam: Configurable marine weather forecast generator[J]. Expert Update , 2003 , 6(3) : 4-10.
- [80] ReiterE, Sripada S, Hunter J, Yu J, & Davy I. Choosing words in computer-generated weather forecasts [J]. Artificial Intelligence , 2005 , 167(1) : 137-169.
- [81] Belz, A. (2008). Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering* , 14(04) , 431-455.
- [82] Belz A, & Kow E. System building cost vs. output quality in data- to- text generation. Proceedings of the 12th European Workshop on Natural Language Generation. Association for Computational Linguistics[C]. 2009 : 16-24.
- [83] Bohnet B, Lareau F, & Wanner L. Automatic production of multilingual environmental information. Proceedings of the 21st Conference on Informatics for Environmental Protection (EnviroInfo-07)[C]. 2007.
- [84] Kukich K. Design of a knowledge- based report generator. Proceedings of the 21st annual meeting on Association for Computational Linguistics[C]. Association for Computational Linguistics , 1983 : 145-150.
- [85] KahnM G, Fagan L M, &Sheiner L B. Combining physiologic models and symbolic methods to interpret time- varying patient data[J]. Methods of information in medicine , 1991 , 30(3) : 167-178.
- [86] Hüske-Kraus D. Suregen-2: A shell system for the generation of clinical documents. Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics[C]. Association for Computational Linguistics , 2003 , 2: 215-218.
- [87] PortetF, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, & Sykes C. Automatic generation of textual summaries from neonatal intensive care data[J]. Artificial Intelligence , 2009 , 173(7) : 789-816.
- [88] B Yao, X Yang, L Lin, M W Lee, and S C Zhu. I2t: image parsing to text description[J]. IEEE Xplore. 2010.
- [89] Y Feng and MLapata. How Many Words Is a Picture Worth? Automatic Caption Generation for News Images. Proc. Assoc. for Computational Linguistics[C]. 2010: 1239-1249.
- [90] Y Feng and M Lapata. Automatic caption generation for news images[J]. IEEE Trans. Pattern Anal. Mach. Intell. , 2013 , 35.
- [91] Y Yang, C LTeo, H Daumé III, and Y Aloimonos. Corpus-guided sentence generation of natural images. EMNLP[C]. 2011.
- [92] G Kulkarni, V Premraj, S Dhar, S Li, Y Choi, A C Berg, and T L Berg. Baby talk: Understanding and

- generating image descriptions. CVPR[C]. 2011.
- [93] Kulkarni, Girish, Premraj, Visruth, Ordonez, Vicente, Dhar, Sag-nik, Li, Siming, Choi, Yejin, Berg, Alexander C, and Berg, Tamara L. Babytalk: Understanding and generating simple im-age descriptions [J]. PAMI, IEEE Transactions on, 2013, 35(12) : 2891-2903.
- [94] Mitchell, Margaret, Han, Xufeng, Dodge, Jesse, Mensch, Alyssa, Goyal, Amit, Berg, Alex, Yamaguchi, Kota, Berg, Tamara, Stratos, Karl, and Daumé III, Hal. Midge: Generating im- age descriptions from computer vision detections. In European Chapter of the Association for Computational Linguistics[C]. ACL, 2012: 747-756.
- [95] Elliott, Desmond and Keller, Frank. Image description using vi-sual dependency representations. EMNLP [C]. 2013 : 1292-1302.
- [96] Hodosh, Micah, Young, Peter, and Hockenmaier, Julia. Framing image description as a ranking task: Data, models and evalua-tion metrics[J]. Journal of Artificial Intelligence Research, 2013 : 853-899.
- [97] A Karpathy and L Fei-Fei. Deep visual-semantic align-ments for generating image descriptions. CVPR[C]. 2015.
- [98] RSocher, A Karpathy, Q V Le, C D Manning, and A Y Ng. Grounded compositional semantics for finding and de-scribing images with sentences. TACL[C]. 2014.
- [99] X Chen and C L Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. CVPR[C]. 2015.
- [100] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang and Alan L Yuille, Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN). ICLR[C]. 2015.
- [101] J Donahue, L A Hendricks, S Guadarrama, M Rohrbach, S Venugopalan, K Saenko, and T Darrell. Long-term recur-rent convolutional networks for visual recognition and de-scription. CVPR[C]. 2015.
- [102] O Vinyals, A Toshev, S Bengio, and D Erhan. Show and tell: A neural image caption generator. CVPR [C]. 2015.
- [103] Xu K, Ba J, Kiros R, Courville A, Salakhutdinov R, Zemel R, & Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. ICML[C]. 2015.
- [104] H Fang, S Gupta, Flandola, R Srivastava, L Deng, P Dollár, J Gao, X He, M Mitchell, J Platt, C L Zitnick, and G Zweig. From captions to visual concepts and back. CVPR[C]. 2015.
- [105] Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, Margaret Mitchell. Language Models for Image Captioning: The Quirks and What Works. arXiv 2015[DB].

作者简介

万小军 博士，北京大学计算机科学技术研究所研究员，博士生导师，主要研究方向为自然语言处理、文本挖掘。Email: wanxiaojun@pku.edu.cn



冯岩松 博士，北京大学计算机科学技术研究所讲师，主要研究方向为自然语言处理。Email：fengyansong@ pku. edu. cn



孙薇薇 博士，北京大学计算机科学技术研究所讲师，主要研究方向为计算语言学。Email：ws@ pku. edu. cn



工业控制计算机研究进展与发展趋势

工业控制计算机专委会

摘要

工业控制计算机是一种采用总线结构，对生产过程及其机电设备、工艺装备进行检测与控制的设备的总称。随着技术的快速发展，目前工业控制计算机的概念已超越了这种传统定义，正在涉及计算机科学中越来越多的技术和领域，向着标准化、网络化、智能化、并行处理和高可靠的方向发展。本文针对工业控制计算机快速发展的现状，紧紧围绕工业控制计算机的发展历程及其关键技术，主要论述了三方面的内容。首先，重点论述了总线技术革新驱动下工业控制计算机的发展历史和现状，并对工业控制计算机发展趋势和前景进行了展望。其次，介绍了当前工业控制领域最重要的嵌入式系统中涌现的新技术，即片上系统、片上网络、片上多核并行处理及多核操作系统等。最后，针对工业控制中高可靠和高安全的特殊领域，介绍了当前广泛应用的容错计算机体系结构、可重构技术以及软件形式化验证技术。

关键词：工业控制计算机，嵌入式系统，容错计算机

Abstract

Industrial control computer is a sort of collectively called devices, which use bus architecture to detect and control production process and its electromechanical devices and technical devices. With fast development of technology, the current concept of industrial control computer is further beyond the traditional definition, involving more and more techniques and fields in computer science. Now, industrial control computer has a development trend of standardization, networking, intelligentization, parallel processing and high reliability. Towards this fast development status of industrial control computer, we mainly discuss three aspects in industrial control computer field in this paper, based on the roadmap and key techniques of industrial control computer. Firstly, the industrial control computer development history and current status driven by bus technique evolution is emphasizedly discussed. And the industrial control computer development trend and perspective is presented. Secondly, new techniques of embedded system, which is very important in current industrial control field, are introduced, such as system on chip, network on chip, chip multi-processing and multi-core operating system. Finally, towards the special field of high reliability and high safety in industrial control, current frequently applicable key techniques are introduced, such as fault tolerance computer architecture, reconfiguration computing technique and software formal verification.

Keywords: industrial control computer, embedded system, fault tolerance computer

1 引言

工业是一个国家国民生产中最重要的组成部分之一，决定着国家的经济命脉和综合

国力。工业的自动化水平代表了一个国家的工业实力，乃至整个综合国力的水平。自从计算机被发明以来，计算机技术不断被运用到工业自动化控制的各个领域中，对工业起到了决定性作用。工业控制计算机也作为一个单独的学科和领域被独立出来，世界各国均投入大量的人力和物力对工业控制计算机进行研究，设计和制造出适用于各个工业领域的工业控制计算机。

一般来说，工业控制计算机是一种采用总线结构，对生产过程及其机电设备、工艺装备进行检测与控制的设备的总称，是用于实现工业生产过程控制和管理的计算机。工业控制计算机是实现工业自动化最重要的设备，涉及的领域和学科非常广泛，几乎涵盖了计算机科学的各个分支。近年来，电子技术和信息技术高速发展，尤其是芯片技术、软件技术、网络技术，以及智能技术的革命性发展，给工业控制计算机发展带来了前所未有的契机，同时也提出了新的挑战。自 20 世纪 70 年代产生以来，工业控制计算机的功能和性能发生了翻天覆地的变化，新技术也层出不穷。本文针对工业控制计算机快速发展的现状，紧紧围绕工业控制计算机的发展历程及其关键技术，主要论述了以下三个方面的内容。

(1) 总线技术革新驱动下的工业控制计算机发展：纵观工业控制计算机的发展历史，工业控制计算机随着微处理器芯片的发明和广泛应用而起源，但其发展更替以总线技术的发展为主线。由于工业控制计算机对标准化、可靠性、可测试性和可维护性的要求，总线技术对工业控制计算机至关重要，工业控制计算机每次历史性更新换代都以总线技术的更替为先导。本文以工业控制现场总线为线索，总结和论述了工业控制计算机发展历史和现状，并对发展趋势进行了展望。

(2) 高度集成及多核并行的工业控制嵌入式处理器与操作系统技术：自从 20 世纪 70 年代产生以来，嵌入式系统最早应用于工业控制的多个领域中，目前已成为工业控制中最为重要、应用最为广泛的计算机系统。多种工业现场控制计算机、自动化设备、工业控制和无线通信终端都是嵌入式系统。近二十年来，嵌入式系统得到了更为高速的发展，由单片机时代发展到高速嵌入式处理器时代，目前已经发展到并行多核片上系统。嵌入式操作系统也随之进入多核并行时代，不仅实现了高实时性、高可靠性，而且实现了片内并行任务调度的高性能处理。

(3) 面向高可靠工业控制领域的容错计算机技术：在工业控制领域，有一类高度可靠、高度安全的应用，这对计算机提出了更为苛刻的要求。这类工业控制计算机需要比普通工业控制计算机具有更高的可靠性和安全性，往往需要进行特别的容错设计。随着电子和信息技术的高速发展，这类工业控制计算机的设计和制造中也不断出现了新的技术和方法。

2 工业控制计算机

2.1 概述

工业控制计算机是工业自动化设备和信息产业基础设施的核心。传统意义上，将用

于工业生产过程的测量、控制和管理的计算机统称为工业控制计算机，简称“工控机”。根据全国工业过程测量和控制标准化技术委员会于 2011 年 7 月 29 日发布的工业控制计算机系统标准定义：工业控制计算机是按常见工业现场条件设计，适用于工业实时检测、监视和控制应用的计算机^[1]。

改革开放以来，工控机为工业自动化、信息产业和国防建设的发展提供了一条低成本的自动化技术方案，促进了国民经济的发展。同时，工控机技术自身也得到了迅速发展。三十多年的发展实践证明：可靠而廉价的工控机适合中国国情。

工控机的主要特点如下：

- 具有标准化、模块化结构，便于组合和扩展；
- 具有系列化、产品化的控制器模块和丰富的 I/O 功能模块，具有广泛的适应性；
- 通用性和开放性强，便于系统集成；
- 恶劣环境条件下适应性强和可靠性高，能长期不间断工作，具有良好的工作稳定性；
- 体积小，成本低，应用系统开发时间短；
- 符合标准化和可互操作性，标准和产品由众多厂家支持和发展。

回顾历史，工控机技术的发展经历了 20 世纪 80 年代的第一代 STD 总线工控机，20 世纪 90 年代的第二代 IPC，21 世纪 00 年代的第三代 CompactPCI 总线工控机以及 21 世纪 10 年代的嵌入式系统时期，而每个时期大约要持续 15 年左右的时间^[19]。STD 总线工控机解决了当时工控机的有无问题；IPC 工控机解决了低成本和 PC 兼容性问题；CompactPCI 总线工控机解决了可靠性和可维护性问题，嵌入式系统解决了小型化、定制化和个性化的问题。目前嵌入式系统应用已经几乎“无所不能、无处不在”；作为第三代工控机技术，CompactPCI 总线工控机主要应用在对坚固性和可靠性有严格要求的电信与生产过程自动化领域，IPC 逐渐由生产过程自动化层转移到管理信息化层，而 STD 总线工控机已经完成其光辉的使命，退出了历史舞台。

2.2 国际研究现状

工业控制计算机技术的起源要追溯到 20 世纪 70 年代。20 世纪 70 年代初期，随着微型计算机技术的兴起，出现了一系列计算机芯片，如 Zilog 公司的 Z80，Motorola 的 68K，Intel 的 8088、8086、MCS51 系列等。这些厂家生产的芯片互不兼容，引脚定义、运行时钟、工作时序都不同，给构造微型计算机及其应用系统造成了极大困难，严重阻碍了微机技术的进步。标准化的问题开始引起了芯片制造商的关注。

标准化的道路有两条。一条是芯片内部总线的标准化，只要所有厂家都按同一种片内总线生产芯片，具有相同的引脚定义和工作时序，就可以实现互连。但一方面，难以找到一条公认的最佳片内总线，另一方面，这与制造商的利益冲突，因此受当时历史条件的局限，难以实现芯片一级标准化。另一条是在模板一级实现标准化。由于模板一级的标准化可以独立于微机芯片，既维护了芯片生产制造商的利益又可以实现产品互连、

协同工作，因此，模板级标准化的方案得以实现和发展。

板级互连总线标准兼顾了大多数芯片制造商的利益，使得各种系列芯片只要经过简单的逻辑转换，就可以与总线标准兼容，具有合理性和先进性，被大众所接受，成为了公认的总线标准，由此也产生了工控机总线标准。工业控制计算机系统的总线是一种传递规定信息的公共通道，通过它可以把各种数据和命令传递到各自要去的地方^[2]。工控机标准总线是工控机系统内部各种独立模板（块）之间进行数据和信息交流的公共通道。按总线标准研制生产的功能模板叫作“OEM”产品，按照标准生产的带有一定数量总线插槽的总线底板叫作“背板”。这样，将各种功能模板插在背板的插槽上就可以构成一个模块化的微型工业计算机系统。

2.2.1 STD 总线工业控制计算机技术的发展

1978 年，美国 Pro-Log 公司和 Mostek 公司发起制定了第一个工业控制计算机总线标准——STD 总线工业 I/O 标准，成立了由数十家企业参加的 STD 总线工控机制造商协会（STDMG）^[4]。STD 总线是业界最早、最具标志性、最具影响力工控机技术标准，具有 56 根信号线，金手指连接方式，支持 8 位 I/O 总线，随后发展成 16 位总线，统称为 STD80；以后进一步发展成 32 位总线，称作 STD32。1987 年 STD 总线标准被国际标准化组织吸收，成为 IEEE961 国际标准。当时，国际上主要的 STD 总线工控机制造商有 Pro-Log、WinSystems、Ziatech 等众多厂商。

采用 STD 总线设计制造的模块化微型计算机系统叫做 STD 总线工业控制计算机。STD 总线工控机是工业型计算机，STD 总线的 16 位数据总线、8MHz 总线频率满足嵌入式和实时性应用要求；特别是它的小板尺寸（4.5 英寸 × 6.5 英寸）、垂直放置无源背板的直插式结构、丰富的工业 I/O OEM 模板、低成本、低功耗、扩展的温度范围、可靠性和良好的可维护性设计，使其在空间和功耗受到严格限制的、可靠性要求较高的工业自动化领域得到了广泛应用。

但是，随着 CPU 的性能和存储器容量的不断扩大，STD 总线的性能越来越不能满足技术升级换代的需要。同时，边缘金手指连接方式在根部引起的断线问题，以及存在系统与外部现场信号之间通过扁平电缆连接的方式造成的接触不良现象，也在一定程度上降低了其可靠性。因此，STD 总线工控机已经淡出工控主流市场。

与 STD 总线处于同一时期并具有一定影响力的还有 Intel 公司于 1977 年推出的用于工业系统的 Multibus 总线，被国际标准化组织批准为 IEEE796 标准。到 1982 年，Multibus 总线模板和系统制造商超过了 100 家。基于 Multibus 总线的典型产品主要有 Sun Microsystems 公司生产的 Sun 1 和 Sun 2 工作站等。

2.2.2 VME 总线工业控制计算机技术的发展

1981 年，针对 Motorola 68K 系列 CPU，Motorola、Mostek、Philip 和 Signetics 公司发明了 VME(Versa Module Eurocard) 总线，从此 VME 总线工业控制机在图像处理、工业控制、实时处理和军事通信中得到了广泛应用。1987 年，VME 总线被批准为国际标准

IEEE1014-1987^[4]。

VME 总线接口为两个 96 芯的针孔连接器 P1 和 P2 (6U 欧洲卡)，数据宽度为 32 位，最大总线速度是 40MB/sec，具有很高的抗震动、抗冲击以及防尘和防腐蚀能力。1996 年的新标准 VME64 (ANSI/VITA1-1994) 将总线数据宽度提升到 64 位，最大数据传输速度为 80MB/s。1997 年，由 Force Computers 制定的 VME64x 总线规范将总线速度提高到 160MB/s^[13]。1999 年，增加了信源同步传输协议功能的 VME320，将总线速度提升到了 320MB/s，但 VME320 并没有被业界广泛采纳。VME 总线工控机是实时控制平台，大多数运行的是实时操作系统，如 UNIX、VxWorks、PSOS、VRTX、PDOS、LynOS 以及 VMEXEC，由操作系统制造商提供专用的软件开发工具开发应用程序。

VME 总线 OEM 产品由众多的制造商支持，主要采用 Motorola 公司的 68K 系列微处理器，如 25M/33M 的 68060，以及 PowerPC 等处理器。现在越来越多的 VME 总线 SBC 的制造商开始采用 Intel 公司和 AMD 公司生产的微处理器，如 DY 4 Systems、Concurrent Technologies、VMIC、PEP Modular Computers，以及 SBS Embedded Computers 等公司。

1987 年 VXI(VME eXtensions for Instrumentation) 总线联合会成立，其目标是定义板级仪器标准，即 VME 扩展仪器仪表总线。VXI 总线工控机以 VME 总线为内核，将欧洲卡尺寸扩展为 A、B、C 和 D 四种。

由于 VME 和 VXI 总线有丰富的控制器与 I/O 功能模板资源，具有“热插拔”和“多主机板并行工作”的技术特点，因此，虽然价格昂贵，但目前还不能完全被 CompactPCI 总线等其他工控机技术取代。在可预见的将来，它们在高可靠性应用领域仍然会占有一席之地。

2.2.3 IPC 技术的发展

IPC(Industrial Personal Computer) 随着 PC 技术的发展而不断发展的。早期的 IPC 是基于 ISA(Industry Standard Architecture) 总线设计的。

1981 年 8 月 12 日 IBM 公司正式推出了基于 8 位 8088 处理器的 IBM PC/XT 机，同时开放了其 8 位/8MHz/62 根信号线的系统总线 PC/XT 总线。1984 年，推出基于 80286 CPU 的 IBM PC/AT，其 16 位数据线/8MHz 频率/98 根信号线的 PC/AT 总线标准被国际标准化组织确定为 ISA 总线标准，即 IEEE P996，被业界广泛接受。随后 PC 借助于规模化的硬件资源、丰富的商业化软件资源和普及化的人才资源，于 20 世纪 80 年代末期开始进军工业控制机市场，形成了风靡全球的以 ISA 总线为基础的 IPC，开创了一个基于 PC 系统的应用时代。

20 世纪 90 年代末期，ISA 总线技术逐渐淘汰，PCI 总线技术开始在 IPC 中占主导地位，使 IPC 通过采用新技术而得以继续发展。因此 IPC 并没有向当时人们预测的那样很快“消亡”，而是得到了更快的发展。

为了满足日益提高的处理器速度对快速传输和处理海量数据的要求，1992 年，Intel 公司定义了 PCI(Peripheral Component Interconnect) 总线作为局部总线，直接用于 CPU 与外设部件，如与显示控制器、网络控制器、内存控制器等互连。为了研制和管理 PCI 标

准，成立了 PCISIG (PCI Special Interest Group)，并先后推出了 PCI Local Bus Specification Rev. 2.1、Rev. 2.2 以及最新版本 Rev. 2.3，PCI 总线也由芯片级互连总线发展成了板级互连总线，成为事实上的 PC 总线国际标准。

由于 PCI 总线的出现，使得支持 PCI 总线的 IPC 得以快速发展。一开始，在一台 IPC 的无源背板上，ISA 和 PCI 插槽共存来相互兼容的，并实现过渡。今天 IPC 上已经没有 ISA 插槽，只有 PCI 以及更先进的其他总线插槽。

PCI 总线的 32 位/33MHz 标准具有 132MB/s 数据传输速度，其 64 位/66MHz 的标准可以达到 528MB/s，所以，相对传统的 ISA 或 VME 总线每秒几兆字节或几十兆字节的传输速度而言，PCI 是高速、高性能总线。

IPC 主要生产企业有 ICS、西门子以及康泰克、研华、凌华、磐仪和艾讯等。

当然，IPC 也存在一定的局限性，限制了其进一步发展和应用：

- 受到机箱结构的限制，散热性能不好，容易引起印制板变形、断线、接触不良等问题，还会造成电子器件寿命缩短以及工作不稳定等；
- 板卡和无源背板之间的金手指边缘接触连接方式，容易造成在振动和冲击过程中瞬间接触不良，引起系统死机；
- 金手指自身在潮湿或腐蚀性气体环境中长期使用，容易氧化或腐蚀，造成系统接触不良，而且多次拔插容易变形；
- 多数板卡通过金属挡片一端固定，在振动力的左右下容易产生微距离逆时针旋转，造成系统信号断路或短路，使系统崩溃；
- 系统机箱表面喷漆处理，不能形成一个完整的导电体，电磁干扰的屏蔽能力和静电释放能力变差；
- 不支持热插拔，故障板卡更换时间长。

因此，从 20 世纪 90 年代开始，IPC 逐渐从对可靠性有特殊要求的电信和流程工业等领域退出。但在可靠性要求不是非常苛刻的工业控制自动化和测试测量自动化以及管理信息化领域，IPC 凭借其优异的品质、低廉的价格、使用和维护的方便性等特点，直到今天仍然占有很大的市场份额。在工业控制自动化发展史上，基于 PC 技术的 IPC，开创了一个绚丽多彩的基于 PC 的时代，对推动工业化和信息化的融合和发展发挥了重要作用，产生了深远的影响。

2.2.4 PC/104 总线嵌入式工业控制计算机技术的发展

在 IPC 发展的时期，嵌入式 PC/104 总线工业控制计算机技术也逐渐兴起。为了满足对空间占用和功率消耗严格限制的嵌入式控制应用的需要，并保持与被广泛接受的 PC 总线架构在软件和硬件上的兼容性，美国 Ampro Computers 公司于 1978 年设计了第一款 PC/104 模块。PC/104 是一种工业计算机总线标准。PC/104 有两个版本：8 位和 16 位，两者分别与 PC 和 PC/AT 相对应。1992 年 2 月，Ampro 联合 12 家公司成立了 PC/104 产业联盟 (PC/104 Consortium)，并于 1992 年 3 月推出了与 ISA 总线兼容的 PC/104 总线标准。1992 年，IEEE 开始着手为 PC 和 PC/AT 总线制定一个精简的 IEEE P996 标准，

PC104 作为基本文件被采纳，叫做 IEEE P996.1 兼容 PC 嵌入式模块标准。1997 年 2 月，该标准扩展为与 PCI 总线兼容的 PC/104-plus 总线标准。

PC/104 总线采用自层叠式结构，不需要背板连接，具有尺寸小、结构紧凑、功耗低、可靠性高等特点，在对于空间、功耗和可靠性要求比较苛刻的嵌入式系统中得到了成功应用。

PC/104 总线主要技术优势如下^[9]：

- 尺寸小、结构紧凑：3.6in × 3.8in (90mm × 96mm) 模块；
- 自层叠：取消了无源背板和机笼，叠加扩展；
- 具有结实、可靠、耐用的针孔连接器：具有恶劣环境适应性、可靠性；
- 板卡四角设安装固定孔：抗震动和冲击；
- PC 完全兼容性：降低开发成本、缩短市场投放时间。

为了满足新兴的速度更快、带宽更高的总线发展的要求，1997 年 PC/104 扩展成 PC/104-Plus，增加了 PCI 总线定义，增强了其生命力。PC/104-Plus 同时支持 ISA 总线和 PCI 总线^[10]。PC/104 总线工控机依靠自身的特点不断地改进完善，一直在其传统优势领域占有一定的市场份额。

2003 年 11 月，PC/104 产业联盟正式发布了 120 引脚的 PCI-104 总线标准。

目前 PC/104 总线工控机的主要制造商有 ADLINK、WinSystems、RTD Embedded Technologies、Fastwel、Advanced Micro Peripherals、ADL Embedded Solutions、Diamond Systems、VersaLogic、TE Connectivity、Eurotech、Sundance Multiprocessor Technology、OpenSystems Media 以及 Samtec 等公司。

2.2.5 CompactPCI 总线工业控制计算机技术的发展

PCI 总线技术的发展、市场的需求以及 IPC 的局限性，促进了新技术的诞生，以 CompactPCI 总线为基础的第三代工控机技术得到了迅速发展和普及。

为了将 PCI 总线应用于嵌入式领域，1994 年成立了 PCI 制造商协会（PICMG），主要动机就是基于 PCI 总线技术为嵌入式计算机（如工业计算机、医疗设备、通信设备、交通设备以及军事系统等）研制通用技术标准。PICMG 总部设在美国，在全球有三个分部，即 PICMG PRC、PICMG Europe 和 PICMG Japan^[8]。

1997 年 8 月，PICMG 发布了第一个 CompactPCI 总线工业控制计算机技术标准 PICMG 2.0 R 1.0，CompactPCI Specification。1997 年 9 月发布了 PICMG 2.0 Rev. 2.1。1999 年 10 月发布了 PICMG 2.0 R 3.0。PICMG 2.0 R3.0 是 CompactPCI 总线的核心标准，也是 CompactPCI 技术的精髓。

CompactPCI 总线主要技术特点如下^[20]：

- 具有 PCI 局部总线的电气特性（PCI Local Bus Specification，PCISIG）；
- 具有工业级欧洲卡封装结构和规格（IEC 60297-3 and-4，Eurocard Specification；IEEE 1101.11，IEEE Standard for Additional Mechanical Specifications for Microcomputers Using IEEE 1101.1 Equipment Practice）；

- 具有 IEC 2mm 坚固的米制针孔连接器 (IEC-61076-4-101, Specification for 2mm Connector Systems);
- 具有电磁辐射屏蔽和静电释放能力 (系统由金属外壳和前、后面板组成的整体导电以及 ESD 电路组成, 表现出良好的电磁兼容性);
- 支持热插拔 (CompactPCI Hot Swap Specification);
- 具有标准的系统管理架构 (PICMG 2.9 R1.0, CompactPCI System Management Specification), 定义了基于 IPMI 接口规范的 IPMB 总线及其管理总线网络。

CompactPCI (以下简称 CPCI) 技术规范使得流行的 PCI 总线兼容结构可以用欧洲卡规格实现。这种模块化的 CPCI 模板可以灵活组合成适合各种不同工业现场应用的系统。为了提高 CPCI 系统的可维护性, 需要在 CPCI 产品上增加热插拔功能, 使得 CPCI 模板可以在不需要切断电源的情况下, 插入或拔出正在运行的系统, 而不影响或破坏系统的正常工作, 从而为高可用性设计奠定基础^[13]。

1998 年 8 月 PICMG 发布了热插拔标准 PICMG 2.1 R1.0, CompactPCI Hot Swap Specification。随后于 2001 年 1 月推出了修订版本 PICMG 2.1 R2.0。2002 年 5 月发布了热插拔基础软件规范 PICMG 2.12 R2.0, Hot Swap Infrastructure Interface Specification。为了控制拔出或插入模板的电气和软件连接过程, CPCI 的连接器插针采用长、中、短三层结构。长针为模板提供预充电电源, 中针连接工作电源和 CPCI 信号, 短针为 BD_SEL# 和 IDSEL# 信号, 用于启动板上电源控制并通知该系统模板所处的状态。

2000 年 2 月, PICMG 发布了具有重要意义的系统管理标准 PICMG 2.9 R1.0, CompactPCI System Management Specification。该标准为 CPCI 系统定义了基于智能平台管理接口 (Intelligent Platform Management Interface, IPMI) 规范的智能平台管理总线 (Intelligent Platform Management Bus, IPMB) 及其管理总线网络。

互为备份的两个专用的系统管理控制器模板 (独立于主 CPU), 实时采集和查询系统所有部件的事件日志, 向系统管理员报告引起系统服务中断的异常事件, 以便及时采取预防措施, 防止系统崩溃。系统所有部件 (包括 CPU 模板、I/O 模板、包交换模板、无源背板、故障切换电路、机箱、电源、冗余磁盘阵列、风扇、网络以及环境条件等) 的监控和管理都统一由管理软件和系统管理控制器通过 IPMI 和 IPMB 网络实现^[14]。

国外 CPCI 产品制造商主要有 Intel、Ziatech、Motorola Computer Group、Kontron、NMS、Force 以及 Performance Technologies 等。

2.2.6 新一代总线技术——PCI Express

21 世纪之初, 随着计算机和通信技术的进一步发展, 新一代的超高速 I/O 接口大量涌现, 比如千兆 (GE) / 万兆 (10GE) 的以太网技术、4Gbit/s/8Gbit/s 的 FC (Fibre Channel) 技术, 使得 PCI 总线的带宽已经不能满足微型计算系统内部大量数据并行读写的要求, PCI 总线也成为系统性能提升的瓶颈, 于是 PCI Express 总线就出现了。

PCI Express 总线是高性能、通用、点到点串行的互连总线, 是面向未来计算平台和电信平台应用设计的, 采用了当今最先进的点到点互连技术、基于交换机的技术和包装协议

(Packetized protocol) 技术，代表了新一代微型计算机总线技术的发展方向和最新动态^[11]。

在 2001 年春季，Intel 公司宣布要用一种新的技术取代 PCI 总线，实现芯片之间的互连，并称之为第三代 I/O 总线技术，即 3GIO (3rd Generation I/O)，并委托 AWG (Arapahoe Work Group) 负责标准的起草工作。2002 年 4 月，3GIO 1.0 规范标准起草完成，递交到包括 Intel、AMD、DELL、IBM 等著名企业参加的 PCISIG 对标准进行审定。2002 年 7 月标准审定结束并对外正式发布，并把 3GIO 更名为 PCI Express。

PCI Express 总线的主要技术优势如下^[11]：

- 双向串行总线，点对点传输，每个传输通道由两对共 4 根低压高速差分信号线组成，提高了抗干扰能力和数据传输速度。单通道单方向的数据传输速率可以达到 2.5Gb/s 以上。
- 支持双向传输模式和数据分通道传输模式。其中数据分通道传输模式，即 PCI Express 总线的 x1、x2、x4、x8、x12、x16 和 x32 多通道连接。
- 支持电源管理、服务质量 (QoS)、可信配置、热插拔或热切换、数据完整性、错误处理等先进功能。
- 与 PCI 总线保持软件的兼容性和一致性，如应用模型、存储结构、软件接口等。
- 降低了系统硬件平台设计的复杂性和难度，降低了系统的开发制造设计成本，提高了系统的性价比和坚固性。

PCI Express 的技术要点是采用了当时流行的点到点串行连接方式，取代了传统的微型计算机总线并行共享架构，第一次实现了从并行到串行的革命性转变。PCI Express 可以理解为是一种将并行 PCI 总线进行打包后变成串行方式在多种传输介质上传输信息的技术，这里的传输介质既可以是布在线路板上的传输线 (trace)，也可以是双绞线 (cable)。PCI Express 技术可以用于板内芯片之间的互连 (chip-to-chip)、模板和模板之间的互连 (board-to-board)，也可以用来实现系统到系统的高速互连 (box-to-box)。

第一代 PCI Express 1.0 的总线数据传输速率 (signaling rate, in bits per second) 为 2.5Gb/s；第二代 PCI Express 2.0 的总线数据传输速率为 5Gb/s。通过 x1、x4、x8、x16 和 x32 连接通道，PCI Express 2.0 总线的传输速率可以达到 80GT/s (GigaTransfer/second)。PCI Express 没有沿用传统共享式总线结构，它采用点对点工作模式 (Peer to Peer, P2P)，每个 PCI 设备都有自己的专用传输线路，这样就无需向整条总线申请带宽，可避免多个设备争抢带宽和总线竞争问题，提高了传输效率和实时性。PCI Express 的主要性能参数见表 1。

表 1 PCI Express 的性能表^[11]

版本号	编码方式	x1 总线			x16 总线	
		传输速率	有效带宽		传输速率	带宽
1. x	8b/10b	2.5GT/s	2Gb/s	250MB/s	40GT/s	4GB/s
2. x	8b/10b	5GT/s	4Gb/s	500MB/s	80GT/s	8GB/s
3. 0	128b/130b	8GT/s	7.877Gb/s	984.6MB/s	128GT/s	15.754GB/s
4. 0	128b/130b	16GT/s	15.754Gb/s	1969.2MB/s	256GT/s	31.508GB/s

从 PCI Express 发展出来的标准目前主要有 SHB (System Host Board) Express、COM (Computer-On-Module) Express、PCI/104-Express、PCIe/104 以及 CompactPCI Express^[17]等。

一块 PCI/104-Express 板卡同时支持 PCI-104 总线和 PCI Express 总线；而 PCIe/104 板卡仅支持 PCI Express 总线。

2.2.7 基于串行总线技术的新一代工业计算机技术的发展

2000 年以来，基于交换式串行互连总线的交换机制技术得到了重视和发展，工控机技术实现了从工业控制向电信和数据通信领域的扩展。

2001 年 9 月，PICMG 将以太网包交换背板总线引入到 CompactPCI 总线标准中，发布了以太网包交换背板标准（CompactPCI Packet Switching Backplane/PSB）PICMG 2.16，将串行以太网在背板上实现，垄断背板设计多年的并行总线技术被打破，使可管理的、冗余的、点到点的、串行包交换互连（switched serial interconnect，也叫 Switched Fabrics）总线技术得到广泛接受。这也为电信语音增值服务设备和基于以太网的工业自动化系统提供了新的技术平台^[4]。

2002 年 12 月 PICMG 发布了基于 Switched Fabrics 技术的更新、更快和功能更强大的开放式平台架构标准 PICMG 3.0，即 AdvancedTCA（Advanced Telecom Computing Architecture，ATCA）。ATCA 比 PICMG 2.16 有更大的规格和容量，更高的背板带宽，对板卡有更严格的管理和控制能力、更高的供电能力以及更强的制冷能力等。ATCA 不是应用在电信上的第一个开放式平台，但它是第一个由电信专家专为电信应用设计的电信平台，主要是为了解决电信系统目前主要面临的系统带宽问题、高可用性问题、现场升级问题、可伸缩性问题、可管理性问题以及可互操作问题，并最终降低成本^[14]。

ATCA 是面向下一代运营通信设备的系列工业标准的总称，它融合了在高速互连技术和下一代微处理器技术中不断提高的可靠性、可管理性和可服务性技术等方面的最新发展成果，满足电信运营对网络设备构建系统（NEBS）、欧洲电信协会标准（ETSI）以及 99.999% 时间可用性的要求。ATCA 还具有模板尺寸大（8U × 280mm × 6HP）、供电能力强（-48V/200W）、可热插拔（hot-swappable）以及支持多协议（Ethernet、InfiniBand、StarFabric、PCI Express、RapidIO）的 Switched Fabrics 技术等特点；支持 Dual Star、Dual Dual Star 以及 Mesh Fabric 拓扑结构；定义了系统管理功能和串行管理接口（Integrated Peripheral Management Interface，IPMI），利用管理软件对模板进行配置，获取状态信息，并能远程关闭故障模板。目前 ATCA 背板的信号传输速度将从现在的 5Gb/s 提高到 10Gb/s，大幅度提升 ATCA 的数据传送能力^[6]。

伴随着 ATCA 技术发展起来的、支持热插拔功能的 AMC 小模块技术，通过简单地更换插入模块，就可以完成 ATCA 系统的现场升级、修改和完善工作，它不仅速度快、费用低，还有效地促进和提升了 ATCA 平台的应用水平^[16]。2005 年 3 月 PICMG 推出了 Advanced Mezzanine Card（AMC）标准；2006 年 7 月，PICMG 发布了最新的技术标准 MicroTCA（Micro Telecommunications Computing Architecture）。MicroTCA 标准定义了 AMC 模块可以直接插入背板上，从而可以构成一种基于交换结构的小型的、价格便宜的但功

能强大的新型系统。MicroTCA 系统包括 AMC 模块、MCH (MicroTCA Carrier Hub)、电源模块 (PM)、制冷单元 (CU)、背板、导轨和机箱等主要部分。AMC 模块有 4 种规格，常用的是 $73.8 \times 13.88 \times 181.5\text{mm}$ ^[15]。一个典型的 19in \times 4U \times 300mm 机箱容纳的 MicroTCA 系统可以支持 10 个 AMC 模块、两个 MCH 和两个 PM。MicroTCA 的背板信号接口采用先进的差分高速串行器和解串器 (Serializer/Deserializer, SerDes) 互连技术设计，半双工带宽可以达到 3.125Gb/s ^[7]。

借助于 AMC 和 MicroTCA 技术，ATCA 在网络设备、存储设备、工业控制、过程控制、测量仪器、军事/国防系统、医疗设备以及其他更广泛的应用领域得到了迅速发展。

2.2.8 基于 PXI 总线总线的模块化仪器仪表架构工业计算机技术的发展

随着 CPCI 技术的发展，一种基于 PXI 总线的模块化虚拟仪器仪表技术出现了。仪器和仪表是工业自动化设备的重要组成部分，工业控制计算机技术在仪器仪表领域的应用形式是虚拟仪器仪表。电子测量仪器的发展大致经历了四代产品：第一代的模拟仪器；第二代的数字化仪器；第三代的智能仪器；第四代的虚拟仪器。在现代电子测量领域中，虚拟仪器无疑是现在和将来自动测试技术发展的主要方向。

区别于传统台式的、用电路实现的、功能固定的模拟仪器，虚拟仪器 (Virtual Instrument) 基于软件技术设计，通过计算机提供的强大图形环境和功能扩展能力，建立图形化的虚拟仪器面板，完成对仪器的控制、数据采集、数据测量和分析以及测量结果显示等功能。它利用工业计算机的标准化、模块化、开放式体系结构，配上相应的仪器驱动软件，使计算机成为一台具有高度测试自动化能力、功能强大的“仪器”。

CompactPCI 向仪器仪表领域的扩展总线就是 PXI (PCI eXtensions for Instrumentation) 总线。PXI 继承了 CompactPCI 总线的坚固的、模块化的欧洲卡机械结构和 PCI 总线电气兼容特性，并增加了专门的同步总线以及机械、电气和软件方面的一些关键性能，定义了用于测试、测量、采集数据、生产制造等应用的完整系统，为测量和自动化系统提供了高性能、高坚固性、低成本的解决方案。

PXI 产生于 1997 年。1998 年美国 National Instruments 公司组织成立了 PXI 系统联盟 (PXI Systems Alliance)，正式推出 PXI 总线标准，使其成为一种开放的工业规范，以满足日益增长的复杂仪器系统需求。PXI 主要是面向“虚拟仪器”市场而设计的，但已经不局限于测试和测量设备，已经迅速向其他工业控制自动化领域扩展，并与 CompactPCI 总线互相补充和融合。PXI 总线工控机不但具有 VXI 的高采样速率、高带宽和高分辨率等特点，而且具有开放性、PC 软件兼容性和低价格等优势。PXI 系统由机箱、系统控制器和外围模块三个基本部分组成。PXI 和 CompactPCI 之间具有互操作性，即在同一个 PXI 系统中可以同时包含 CompactPCI 和 PXI 模块，而不产生任何冲突^[12]。

2005 年年底，PXI 中引入了 PXI Express 技术，提高了其总线带宽，以满足更多的应用需求。利用 PCI Express 技术，PXI Express 将 PXI 中的可用带宽提高了 45 倍多，即从 132MB/s 提高到 6GB/s ；与此同时，这还可以维持与 PXI 模块间的软件、硬件兼容性，使 PXI 可以用于很多新型应用领域，其中很多领域在以前只能由昂贵的专用硬件实现^[18]。

PXI 有 3U 和 6U 两种规格。一般来说，3U PXI 产品用于构造便携式或小型 ATE 测试设备、数据采集系统、监控系统以及其他工业自动化系统。6U PXI 产品主要应用在高密度、高性能和大型 ATE 设备中。

目前，PXI 已经成为测试和测量领域继通用接口总线（GPIB，IEEE488）之后成长最快的标准技术。无论是生产厂家的数量，产品的种类和数量，还是系统应用的数量都有大幅度增长，越来越多的项目转向 PXI 解决方案。其中包括应用于测量的 CPCI 模板在内，现在可用于 PXI 系统集成的模板已经超过 1500 种，产品相当丰富。产品的性能也得到了显著的提升，如数据采样速度已经达到 2GHz，测量精度提升到 7 位数字，射频测量带宽也达到了 3GHz。而且 PXI 的性能（如测量速度和精度等）还在不断地提高。可以预计在不久的将来，PXI 可能超过甚至取代传统盒式测量仪器，占据中、低频段的高精度测量设备市场的主要份额。

中国台湾和国际上主要的 PXI 产品制造商有 National Instruments（美国国家仪器公司）、ADLINK Technology（中国台湾凌华科技）、CHROMA ATE（中国台湾中茂电子）、Aeroflex（美国艾法斯）等。

2.3 国内研究进展

为了把国外工控机先进技术引入到国内来，促进国内工控机技术的健康发展，推动国家四个现代化建设，国内工控领域的著名专家、学科带头人、技术骨干和高级管理人员发起创意，并于 1984 年 6 月 11 日成立中国电子学会电子计算机分会工业计算机学组，1986 年 3 月 22 日正式成立了中国计算机学会工业控制计算机专委会。第一届～第二届专委会主任是陈令，秘书长是唐怀斌；第三届专委会主任是魏庆福，秘书长是段明祥；第四届～第五届专委会主任是杨孟飞，秘书长是刘鑫；第六届专委会主任是刘鑫，秘书长是杨桦。

工控机专委会成立以后，积极开展工控领域的学术交流、战略研究、标准制定、专业培训等，提高科研、教学、应用水平，促进研究成果的应用和转化，推动了国内工控领域的技术进步，促进了 STD 总线工控机、IPC 以及 CompactPCI 总线工控机技术在中国的研制、生产、推广和应用。

20 世纪 70 年代，随着计算机技术的飞速发展，将计算机技术应用到工业生产领域，提高产品的产量和品质成了当时工业控制行业所追求的目标。但国内微机技术和电子集成电路技术比较落后，国内还没有自己生产的相关计算机产品可以应用。1978 年美国 Pro-Log 公司发明了 STD 总线，推出了 STD 工控机，在国内工控界引起了强烈反响。

1983 年年底，由中国香港新华社科技组牵头，中国香港耀阳高科技有限公司总裁吴江先生和北京康拓公司的魏庆福先生率领一行 9 人的工作小组前往美国蒙纳利市的 Pro-Log 公司进行访问、调研。他们不仅获得了一些珍贵的 STD 总线技术和产品资料，还带回了大量 STD 总线模块产品。STD 总线兼容 PC 总线，其小板结构既符合计算机的集成化发展方向，又适合工业现场的应用。工作组成员相信如在中国推广 STD 总线，将能改变我国微型计算机工业落后的面貌。

在工作组成员的积极倡导与宣传下，国内工业控制届科研人员严格遵循国际 STDMG 的标准和规范，对 Pro-Log 的技术进行消化吸收和仿制。1988 年成立了中国 STD 工控机制造商协会（STDMG/PRC），开始了国内自主研制工业控制计算机的历史。中国 STD 总线工控机融合了 STD 总线技术、Intel CPU 技术以及 Microsoft 的 DOS/Windows 技术，走出了一条中国特色的发展道路，取得了辉煌的成就。由于国产 STD 总线工控机研制的巨大成功，完全垄断了国内市场，将国外的 STD 总线工控机产品挡在了国门之外。

1995 年，为了通过与 Intel 公司的合作，进一步推动国内工控机技术的发展，康拓工业电脑公司总经理魏庆福，率领张庆汉、刘鑫和韦红文一行 4 人，到美国 Intel 公司总部凤凰城（Pheonix, Arizona）考察，并将 Intel 当时刚刚推出的先进 Flash Memory 技术^[21] 和嵌入式 80386EX 技术带回国内，使 STD 总线工控机技术的发展上了一个新的台阶。当时比较成功的产品是增强型 V40 系统Ⅱ^[22]、V20 系统以及采用 All-in-One 技术的 386EX 系统^[23] 和 486SX/DX 系统等。主要制造商有北京康拓公司、四通公司工控部、北京工业大学电子厂、北京华胜工控工程公司、北京华远自动化系统有限公司、北京工业控制计算机厂以及重庆工业自动化仪表研究所等。

STD 总线工控机采用机笼式安装结构，具有标准化、开放式、模块化、组合化、尺寸小、成本低、兼容 PC 等特点，并且设计、开发、调试简单，得到了当时急需用廉价而可靠的计算机来改造和提升传统产业的中小企业的广泛欢迎和采用。实践证明，STD 总线适合中国国情，STD 工控机成为当时我国工业控制领域中的一种主导产品，国内的总安装容量接近 20 万套，在中国工控机发展史上留下了辉煌的一页。

IPC 在中国的发展大致可以分为三个阶段。第一阶段是从 20 世纪 80 年代末到 90 年代初，这时市场上主要是国外品牌的昂贵产品。第二阶段是从 1991 年到 1996，中国台湾生产的价位适中的 IPC 开始大量进入中国大陆市场，这在很大程度上加速了 IPC 市场的发展，IPC 的应用也从传统工业控制向数据通信、电信、电力等对可靠性要求较高的行业延伸。第三阶段是从 1997 年开始的，中国大陆本土的 IPC 厂商开始进入该市场，促使 IPC 的价格不断降低，也使工控机的应用水平和应用行业发生极大变化，应用范围不断扩大，IPC 也随之发展成了中国第二代主流工控机技术^[4]。

1993 年，鉴于国产 IPC 进展缓慢，开发多限于外围卡，主板开发处于空白阶段，致使境外产品大量拥入，占去了大部分市场的状况，中国计算机学会工业控制计算机专业委员会在国家科委、机械工业部、航天工业总公司、全国电子信息推广办、北京电子振兴办以及中国计算机学会等部委、学术团体的大力支持下，及时召开了“工业 PC 国产化研讨会”。各界就组织起来，抓住机遇，发挥各自优势，形成群体力量，坚定地走取长补短、减少重复、轻型结构、优势集成的全新的联合开发道路达成共识。为落实会议的成果和计划，由北京康拓工业电脑公司、北京华胜工控工程公司、北京华远自动化系统有限公司、北京工业大学电子厂、北京工业控制计算机厂、重庆工业自动化仪表研究所等全国近 20 家工控界精英单位组成了“全国工业 PC 联合开发委员会”^[4]。

1994 年，“联合开发委员会”推出了“IPC8500 系列工业控制计算机”，并正式通过了部级设计定型鉴定。以北京工业大学电子集团工控所董英斌教授为项目主持人的课题组，研制了 All-in-One 结构的 IPC386SX/486SLC 平台系统主机板；北京康拓工业电脑公

司、北京工控机厂研制生产了 IPC8500 系列工业控制机分体式 19 英寸标准机箱。

从技术上讲，“全国工业 PC 联合开发委员会”推出了 IPC8500 系列工业控制机产品，开发有自主版权的组态软件，形成更高层次的开放式系统产品，使 IPC 在我国中小企业自动化甚至管理控制一体化上发挥了重要作用，逐渐占领了国内市场。

目前，国内 IPC 的技术水平与国际上是同步的，IPC 的大小品牌有 15 个左右，主要有深圳研祥的 EVOC、华北工控的 NORCO、北京康拓公司的 IPC8500 等。在国内 IPC 产品中，深圳英德斯科技公司的产品设计具有一定的代表性，全部采用宽电压输入，满足大部分恶劣现场的供电要求，以保证实际使用时不会因电压波动导致系统出现故障；所有接口都有高防护性，可以防护一定强度的接触静电冲击和非接触静电冲击；产品在无风扇情况下高温工作范围宽；可维护行好，维修时间短；采用快速导热的方式设计整机的散热，以保证在高温聚热的情况下，系统能快速导热；软件的可移植性好，能够满足各种操作系统的运行，如 Windows、Linux、安卓、中标麒麟等。

PC/104 作为针对嵌入式控制而设计的工业控制计算机，在国内的发展基本上是与国际同步的，国产品牌几乎垄断了国内的嵌入式市场，主要应用在电力、铁路、轻轨等具有高风险性且要求高可靠性的领域。主要制造商有盛博科技、英贝特、华北工控等。其中盛博科技公司生产的 PC/104 产品是国产品牌中的佼佼者，产品已经走出国门，远销到了欧洲市场。

作为新一代主流工控机技术，CompactPCI 工控机标准于 1997 年发布之初就备受业界瞩目。相对于以往的 STD 和 IPC，它具有开放性、良好的散热性、高稳定性、高可靠性及可热插拔等特点，非常适合于工业现场和信息产业基础设备的应用，被众多业内人士认为是继 STD 和 IPC 之后的第三代工控机的技术标准^[3]。采用模块化的 CompactPCI 总线工控机技术开发产品，可以缩短开发时间、降低设计费用、降低维护费用、提升系统的整体性能。“CompactPCI 是 PCI 总线的电气和软件加上欧洲卡，它具有在不关闭系统的情况下‘即插即用’功能，该功能的实现对高可用系统和容错系统非常重要”，2004 年度科技部科技型中小企业技术创新基金项目指南中的这段话，概括出了 CompactPCI 总线工控机的主要特点和重要性。国家“发改委”2004 年也将 CompactPCI 总线工控机列为主要产业化项目之一^[3]。

1998 年 2 月 20 日，为了更好地将国际先进 CompactPCI 工业计算机技术引入国内，推动国内工业控制计算机技术的同步发展，工业和信息化部成立了中国计算机行业协会 PICMG/PRC（工业计算机专业委员会），同时也是国际 PICMG 协会的执行委员（PICMG China）。PICMG/PRC 的宗旨是，通过与国际 PICMG 协会的深入合作，在中国携手大专院校和工控企业，联合发展开放式标准，将 PICMG 的新技术技术用于工业和电信计算机系统的研制，推动我国工业计算机和自动化产业快速发展。PICMG/PRC 从 2002 年组织“第一届 PICMG 技术年会”开始，每年组织一次技术年会，至今已经连续举办了十三届技术年会；每年举办 3~4 次技术培训；出版了《PICMG/PRC 协会通讯》；为国内工控行业和电信行业了解、接受、掌握和应用 CompactPCI 和 AdvancedTCA 技术发挥了不可替代的作用，也为企业培养了数以千计的专业技术人员。

在 PICMG/PRC 的强力推动下，CPCI 产品在国内发展很快，有自主产品的厂家已经

发展到数十家，产品面广。CompactPCI 工业计算机制造商主要有北京康拓科技有限公司、北京轩宇空间科技有限公司、深圳研祥、中国船舶重工集团第七零九所、七一六所、中国航天科工集团七零六所、航天西安七七一所、电子部上海 32 所等，它们可以研制和生产具有核心技术的系统平台产品。内地的其他单位，如北京新松佳和、北京兴夏机电、深圳研祥、上海鼎钛克、北京理工大学、北京寰龙科技、北京瑞赛科技、北京世纪兴元、北京旋极信息技术、北京中科泛华、陕西海泰、四川纵横仪器公司、北京方天科技等，主要提供各具特色的配套 I/O 产品。在电信产业方面，CompactPCI 相关产品供应商主要有东进技术、杭州三汇、北京林克海德、上海倍生电子科技、杭州迈可行、泰信科技、新太科技、炎黄新星等。电信巨头华为、中兴、大唐等公司也基于 CompactPCI 相关技术标准自主研制了一系列新产品，并开始打入国际市场。但总的来看，中国自主研制的 CompactPCI 产品在市场上占据了主导地位。

作为一个融合通信及数据网络应用的、高性价比的、基于模块化结构的、兼容的并可扩展的先进电信计算平台，AdvancedTCA 在国内的发展与国际上是同步的。北京方天科技、深圳研祥智能、上海恒为科技、华为、中兴、大唐等公司都研制了一系列 ATCA 产品，如高性能的 ATCA 刀片服务器等。全球领先的信息与通信解决方案供应商华为公司已经基于 ATCA 标准设计了 70 余种产品。中兴于 2005 年开始了 ATCA 的研制工作，推出了一系列刀片服务器产品以及基于 ATCA 架构的 ZXIA1000 平台。采用 ATCA 技术设计的 ZXIA1000 平台，具有高性能、高可靠性、高可扩展性、高可维护性，提高了单位面积的处理密度，并大大降低了单位处理能力所消耗的能源，已经分别在浙江移动与江苏移动的彩信和 WAP 网关上顺利应用。

国内 PXI 总线模块化工业计算机的发展相对比较缓慢，产品主要被美国和中国台湾地区的产品垄断。北京中科泛华测控技术有限公司、北京航天测控技术有限公司、泛华恒兴、北京康拓科技有限公司以及北京轩宇空间科技有限公司是国内研制 PXI 总线工业计算机的主要代表者。

2.4 国内外研究进展比较

综上所述，模块化工业控制计算机的发展是伴随着板级互连总线的发展而发展的；板级互连总线又是从芯片级互连总线演变而来的。无论是 STD 总线标准、ISA 总线标准、PCI 总线标准，还是今天的 AdvancedTCA 总线标准，都是由西方国际标准化组织或行业协会制定的，国内企业只是参与者，还不是主导者。在中国计算机学会工业控制计算机专业委员会的组织和指导下，我国工业控制计算机技术的发展也经历了从落后，到同步发展，到部分赶超的过程，发展势头强劲。国产 STD 总线工业控制计算机，虽然在国内起步晚，但后来居上，以其“All-in-One”的先进设计思想、“总线复用周期窃取”设计技术和运行的稳定性，垄断了国内市场，使欧美的产品完全退出中国。嵌入式 PC/104 总线工控机走出国门，远销欧洲。CompactPCI 总线工业计算机和 AdvancedTCA 工业计算机的发展水平是与国际同步的，华为和中兴公司设计的相关产品，处于国际领先水平，而且产品伴随着其通信和网络产品，远销到世界各地。

2.5 发展趋势与展望

随着芯片和模板的密度越来越大，速度越来越快，传统的并行总线逐渐成为系统性能提高的主要瓶颈。与此同时，串行总线的性能也在不断地提高，这也进一步推动了计算机的共享并行总线技术向高速、独占、点到点的串行总线技术方向转移^[4]。因此，2011年，PICMG推出了相应的技术标准CompactPCI Serial，即PICMG CPCI-S.0，解决了模块化工业计算机内部模板之间的串行高速互连问题，实现了海量数据的高速传输，代表了未来工业计算机平台技术的发展方向。CompactPCI Serial采用串行PCI Express、SATA、以太网以及USB，扩展了PCI总线的性能，支持热切换(Hot Swap)功能，可以应用于机器人、机器控制、工业自动化、计算机语音电话和电信、医疗设备、交通运输以及航空航天等国防领域^[5]，具有广泛的应用前景。

随着电子信息技术的高速发展，尤其是芯片技术、软件技术、网络技术以及智能技术取得突破性和日新月异的进展，给工业控制计算机技术的发展带来了前所未有的机遇和挑战。互连网络以及各种智能终端的普及，不仅使人们的生活发生了翻天覆地的变化，同时还引起了工业控制领域的一场革命。伴随着信息产业的发展以及互联网、移动互联网和物联网的三网深层次融合，信息安全问题日益突出。如何研制具有功能安全和信息安全功能的安全型工业计算机，以确保信息系统的安全甚至国家安全，是所有工业计算机行业从业人员的崇高责任和艰巨任务，也是未来工业控制计算机的一个重要发展方向。

未来，信息生产力将取代工业生产力占据社会发展的主导地位。在信息生产力的推动下，使工业从初级的自动化、中级的工业化与信息化的高层次融合，发展到智能化阶段，即信息生产力成为一种最具活力的、先进的生产力形态，具备高度智能化、网络化、高渗透性和全球化运行能力。在这个信息革命时代，工业控制计算机技术必将会发挥不可替代的关键作用，工业计算机产业正面临前所未有的良好发展机遇。

3 嵌入式系统

3.1 概述

嵌入式系统是工业控制计算机的重要组成部分。早在20世纪70年代，人类首次研制出微处理器芯片。伴随微处理器的产生，嵌入式系统蓬勃发展，其最先应用于工业控制领域，也可以说嵌入式系统是为工业控制而产生的。目前，嵌入式系统已经深入到我们日常生活的各个领域，但是工业控制仍然是嵌入式系统最为重要的研究和应用领域，嵌入式计算机系统也是工业控制中最为重要、应用最为广泛的计算机系统。

嵌入式系统一般以应用为中心，以计算机技术为基础，软硬件可裁剪，适用于应用系统对功能、可靠性、成本、体积、功耗等严格要求的专用计算机系统。IEEE 给出的嵌入式系统标准定义是“用于控制、监控或者辅助操作机器和设备的装置”（Devices used to control, monitor, or assist the operation of equipment, machinery or plants）。实际上，嵌入式系统本身是一个外延极广的名词，凡是与产品结合在一起的具有嵌入式特点的控制系统都可以叫嵌入式系统，有时很难以给它下一个准确的定义。

20世纪70年代出现的单片机是嵌入式系统的雏形，单片机的出现使得汽车、家电、工业设备、通信装置等产品可以通过内嵌电子装置来获得更好的使用性能。从20世纪80年代早期开始，开始基于商业级“操作系统”编写嵌入式应用软件，缩短开发周期，降低开发费用和提高开发效率，“嵌入式系统”的概念也在这时逐渐形成了。从2009年开始，由于物联网和云计算的兴起，嵌入式系统得到了极大的发展，成为国内外的热点。21世纪10年代，随着数字终端与移动终端的快速发展和迅速普及，嵌入式系统进入到前所未有的高速发展时期。目前嵌入式系统的重要发展趋势是集成度越来越高，功耗却越来越低，成本也越来越低，其热点集中在系统级集成解决方案片上系统（SoC）、多核并行处理技术、片上网络（NoC）、三维立体封装技术和多核并行操作系统等上。

3.2 国际研究现状

嵌入式系统主要包括嵌入式处理器和嵌入式实时操作系统两个大的领域。当前，嵌入式处理器不断提供更高的集成度、更强的处理能力和更低的功耗。为了实现这些目标，嵌入式处理器的设计思路体现在两个主要的方向。一是朝着片上系统的方向发展，把更多的处理器周边电路、各种总线控制器，以及I/O都集中在一个芯片内部，实现更高的集成度，甚至提出了SIP技术，把多个芯片封装在一起，形成一个三维的集成电路系统（包括处理器、存储器和外设），一个芯片模块实现一台计算机的功能。另一个方向是朝着多处理器内核的方向发展，这是为了解决不断提高处理器主频带来的瓶颈问题，多个处理器内核集成在一个芯片内部，不仅提高了集成度，而且提供了更强的处理能力，并把功耗维持在一个较低的水平。甚至把两个方向集中在一起，形成多核的片上系统。

3.2.1 片上系统

片上系统（System on Chip, SoC）技术是集成电路工艺不断发展的产物，随着深亚微米技术的出现，使得计算机系统能够集成到一个芯片上，SoC将处理器、存储器、各种接口电路、射频电路、模拟电路、传感器甚至微机电系统（Micro Electro Mechanical System, MEMS）等集成在单一芯片上，以实现一个完整系统的功能。这种技术可以显著提高电路的集成度和可靠性，降低体积和功耗，因而成为实现电子系统微小型化、高性能和高可靠性的有效手段。SoC技术是在ASIC技术的基础上融合了IP核等一系列技术发展起来的，自20世纪90年代后期出现以来，受到了学术界和工业界的极大关注。

1994年Motorola公司发布了Flex Core系统，1995年LSI Logic公司采用了SoC技术

进行产品设计，这可以说是基于 IP 的 SoC 设计的最早先例。2000 年美国的 Altera 公司提出了基于可编程逻辑器件的 SoC 解决方案，这种 SoC 又称为 SoPC(System on Programmable Chip)。而美国的 Xilinx 公司从 1985 年至今，不断推出新的可编程器件和工具，对 SoC 的发展起到了重要的推动作用。在 2013 年 10 月左右，处理器业界的龙头公司 Intel 宣布了以“夸克”(Quark) 命名的新平台，主要面向嵌入式智能设备以及可穿戴设备，尤其是物联网领域。夸克是目前已知组成物质的最小基本粒子，Intel 借用这个名字也可谓恰如其分，“夸克”无论面积还是功耗都要小得多。最重要的是，夸克采用了 SoC 设计，其大小相当于现有低端 Atom 的大概十分之一到五分之一，采用单芯片设计，32nm 工艺制造，内核面积应该不到 10mm^2 ，封装尺寸为 $15\text{mm} \times 15\text{mm}$ ，功耗不超过 100mW 。以首发的 Quark X1000 为例，其采用 32 位架构，单内核单线程，主频 400MHz ，辅助缓存 16KB ，内存支持单通道 DDR3-800 2GB，扩展支持两条 PCI-E 2.0，输入/输出支持三个 USB 2.0、16 个 GPIO、两个 UART、两个 LAN，热设计功耗仅仅 2.2W 。

在军事、航空和航天领域，各国的机构也较早地开始了 SoC 的相关技术研究。以美国 NASA 为例，1997 年它提出一项 X2000 计划（即深空系统技术计划），以开发以微电子技术为中心的系统技术，实现以“较快、较好、较省”的空间技术发展方针为目的的空间电子技术发展规划，并确定了以 SoC 为重要的发展方向。

从技术发展角度，SoC 可分为三类：一类是 CSoC(Configurable SoC)，以学术研究机构为主导，注重体系结构探索性工作；另一类是 SoPC(System on Programmable Chip)，以 FPGA 厂商和科研机构为主导，适合多品种少批量产品开发；第三类是 ASIC SoC，以微处理器（跨国巨头）和芯片设计公司为主导，追求良好的性价比，适合大批量规模生产。

在学术界，国际上有专门的 SoC 会议：IEEE SoC Conference (SOCC) 和 IEEE International Conference on VLSI-SoC(VLSI-SOC)。此外众多关于 FPGA 以及低功耗的会议都有大量 SoC 方面的文章参会交流，例如：ACM/SIGDA International Symposium on Field Programmable Gate Arrays 是 FPGA 领域最全面的顶级会议，IEEE International Conference on Field Programmable Logic and Applications (FPL) 是 FPGA 方面最早的会议，IEEE International Symposium on Field Programmable Custom Computing Machines(FCCM) 是 FPGA 应用方面的重要会议，IEEE International Symposium on Low Power Electronics and Design (ISLPED) 是低功耗方面的顶级会议。

3.2.2 多核处理器

单芯片多处理器 (Chip Multi-Processor, CMP)，简称多核处理器。这种技术是随着处理器主频的不断提高，工作频率已经接近极限，由此带来功耗、散热等诸多问题而产生的。在单纯提高主频的情况下，影响了处理器的稳定性和可靠性。仅通过提高时钟频率已经不能提高处理器的性能，因此多核处理器成为处理器发展的趋势，尤其在工业控制等对功耗、稳定性和可靠性要求严格的领域，多核处理器已成为高性能处理器发展的主流趋势。

多核处理器将多个计算内核集成在一个处理器芯片中，从而提高计算能力，其概念

早在 1996 年由美国斯坦福大学提出，当时被称为单芯片多处理器体系结构，也就是通常所说的多核处理器技术，当内核的数目比较多的时候，称为“众核处理器”。但由于当时芯片制造技术正在迅速提高，处理器主频按照摩尔定律不断刷新，制造工艺以及功耗限制带来的矛盾并不突出，因此这种并行技术并没有得到足够的重视和发展。随着芯片主频的提高，制造工艺以及功耗的问题逐渐突出，因此多核处理器逐渐成为主流。由于多核处理器采用了相对简单的单线程微处理器作为处理器内核，使其具有以下优点。

- 高主频：多核处理器中绝大部分信号局限于处理器内核内，包含极少的全局信号，因此线延迟对其影响比较小，可以实现高主频。
- 设计和验证周期短：微处理器厂商一般采用现有的成熟单核处理器作为处理器内核，从而可缩短设计和验证周期，节省研发成本。
- 控制逻辑简单，扩展性好，易于实现。
- 功耗低：通过动态调节电压/频率、负载优化分布等，可有效降低功耗。
- 通信延迟低：多核处理器采用共享缓存或者内存的方式，多线程的通信延迟较低。

多处理系统的架构分为对称多处理（Symmetric Multi-Processing, SMP）结构和非对称多处理（Asymmetric Multi-Processing, AMP）结构两类^[24]。一般而言，同构多核结构的处理器适用于通用处理器，而异构多核结构的处理器适用于各种专用的应用场合。目前市场上的主流多核处理器，例如 Intel、AMD、PowerPC 和 ARM 处理器，都是同构的多核处理器，其适用范围比较广泛，可用于通用的处理和计算。异构多核处理器的结构呈现多样化，一般多应用于专用领域。例如 IBM 推出的 Cell 处理器是一种典型的异构架构。应用于高可靠领域的双核异构处理器包括一个通用处理器内核和一个专用 Java 处理器内核^[25]。此外，一种采用通用处理器内核加 DSP 内核的双核设计模式是一种较为通用的异构多核结构。

早在 2005 年，Intel 和 AMD 的新型处理器融入了多核处理器结构。新安腾处理器的开发代码为 Montecito，采用双核设计，拥有最少 18MB 片内缓存，采取 90nm 工艺制造，它的设计称得上是对当时芯片业的挑战。2005 年 4 月，Intel 推出了第一款供个人使用的双核处理器，打开了处理器历史新的一页。2006 年秋，Intel 公司发布了集成有 80 个 CPU 内核的处理器，用于技术研发，每个 CPU 内核中有两个浮点运算器。利用这种芯片，Intel 正在继续研究集成多个 CPU 内核的处理器内部架构。

2007 年 8 月，美国 Tilera 公司推出面向嵌入式设备的 64 核处理器芯片 TILE64，其中集成了 64 个网格状排列的 3 路 VLIW CPU 内核，每个内核都是完整的处理器。TILE64 包括 5MB 片上缓存和两组板上 256MB DDR2 SDRAM，工作频率为 500MHz ~ 866MHz，运算速度达到 443 亿次/秒，功耗为 15 ~ 22W，I/O 带宽高达 50Gb/s，并提供多种以太网接口。

Boeing 公司的 OPERA(The On-board Processing Expandable Reconfigurable Architecture) 项目致力于研究空间应用处理器。OPERA 采用的芯片为带有抗辐射加固设计 (RHBD) 的 49 核处理芯片 Maestro，它是 Tilera 公司的 TILE64 处理器的抗辐射加固版本，运行速

度为 450 亿/秒，时钟频率为 310MHz，在该时钟频率下浮点运算速度达到 22 亿次/秒。

在 ARM 的众多处理器系列中，基于 Cortex A5、A8、A9 和 A15 内核进行了多核版本的开发^[26]。以苹果公司数据终端所用处理器为例，以 ARM 体系结构为基础，iPhone 5s 采用 A7 双核 64 位处理器，采用了三星 28nm 制程的 Hi K 金属栅极技术，包含 10 亿多个晶体管，内核（Die）大小为 102mm²；而 iPhone 6 则采用了 A8 四核 64 位处理器，采用更为先进的 20nm 工艺。

作为最重要的处理器体系结构，PowerPC 体系结构的处理器形成了庞大的系列，在通信、工控、航天国防等要求高性能和高可靠性的领域得到广泛应用。在面向服务器的 Power 系列处理器中，从 1993 年发布的 Power 2 开始就采用了多核技术，每个芯片封装了 1 500 万个晶体管。被称为 P2SC（Power 2 Scalable Chip）的超级芯片使用 CMOS-6S 技术，用一个芯片实现了 Power2 8 个内核的架构。使用这种处理器的 32 个节点的 DEEP BLUE（深蓝）超级计算机，在 1997 年战胜了国际象棋冠军卡斯帕罗夫。发布于 2001 年的 Power 4 每个芯片封装了 1.74 亿个晶体管，每个处理器包含两个 64 位 1GHz + Power PC 内核。此外，Power 4 支持分区技术，可以将芯片切分成多个单元，运行于不同的操作系统上。Power 5 利用了更快的片内通信技术、芯片多处理技术、同时多线程（Simultaneous MultiThreading, SMT）技术，支持微分区（Advanced Virtualization）功能，可将一个处理器内核虚拟切分成多个处理器，最小的分配粒度为 0.1 个 CPU，共享使用粒度达 1/100 个 CPU。Power 6 通过 ViVA-2（Virtual Vector Architecture）技术提高了向量处理性能，最多支持 1 024 个虚拟分区（Power 5 最多支持 256 个分区）。Power PC 系列是 1993 年从 Power 架构发展出来的一个分支，该系列更专注于嵌入式处理系统，其架构的特点是开放、内核非常小，这样可在同一芯片上集成更多功能。该系列主要包括 Power PC 900、Power PC 700、Power PC 600 和 Power PC 400 系列，其中 Power PC 400 系列最终发展为 PowerPC 405、PowerPC 440、PowerPC 460 3 大系列。目前 Power PC 476 处理器已经开始支持多核处理技术，可见多核已经成为嵌入式系统的一种重要趋势。

随着芯片集成度越来越高，设备越做越小，传统的单核处理器结构越来越不能满足工业控制领域呈指数级增长的计算规模要求，目前片上系统和多核有融合之势。目前，国际上开始有多家公司建立多处理器片上系统（Multi-Processor System on Chip, MPSoC）平台。如 STI（Sony、Toshiba 和 IBM）的 Cell 九核处理器、ARM 公司的 MPCore 四核处理器等。

3.2.3 片上网络

随着电路集成度的提高，SoC 在发展过程中遭遇瓶颈，研究人员借鉴计算机网络和并行计算技术，提出了片上网络（Network on Chip, NoC）的概念，目前成为一个新的研究和应用热点。NoC 的概念最初在 2000 年左右被学术界提出。NoC 可理解成是基于网络通信的多核 SoC（MPSoC），通过路由和报文交换技术完成通信任务。与 MPSoC 相比较，NoC 关注的是通信方式，而 MPSoC 更多地强调多核。从概念上讲，NoC 只是 MPSoC 的一个分支，但 NoC 被广泛认为是未来多/众核系统的发展方向，这种众核处理体系结构能够

为工业控制计算机提供更为强大的功能和性能支持。

2007 年在美国普林斯顿大学由 IEEE 专门组织召开 NoC 国际研讨会 (IEEE International Symposium on Networks on Chip)，标志着 NoC 已成为一个相对独立的研究领域。据初步统计，国际上已有超过 90 多家研究机构对 NoC 各个方面的问题进行着研究，研究范围非常广泛，从体系结构到物理实现，包括系统仿真及综合、路由策略、任务映射、通信协议、功耗管理、应用服务、服务质量、设计与验证方法、工具等诸多方面。国外比较著名的研究成果有麻省理工学院 (MIT) 的 RAW、瑞典皇家理工学院 (KTH) 的 Nostrum、斯坦福大学的 Xpipes、英国曼彻斯特大学的 CHAIN，以及 Intel 公司的 Teraflops 处理器、Philips 公司的 ASEthereal、Boeing 公司的 TILE64 处理器等。

RAW 是由 MIT 开发，包含 16 个 Tile 处理器内核多核处理器。它采用 4×4 的网状拓扑，时钟频率可达 425MHz，共包含 4 个 32 位全双工的片上网络，其中两个为静态网络，两个为动态网络。静态网络通信模型在编译时确定。两个动态网络一个用于存储器的读写，另一个负责与开发者的交互。Nostrum 是 KTH 面向多媒体应用领域开发的 NoC 体系结构研究及设计平台，采用规则的 2D 网格拓扑结构，每个路由器与另外四个路由器相连，每个 IP 核与一个路由器相连，网络采用热土豆路由算法，支持组播路由，并提供 GT (Guaranteed Through put service)、BE (Best Effort service) 两种通信服务。CHAIN (Chip Area Interconnect) 是曼彻斯特大学研发的完全使用异步握手信号进行数据电路交换的自定义 NoC，采用细粒度的流水线链路，通过将发送信息的交换网络和响应信息的交换网络分离，降低了源节点和目的节点的耦合度，网络采用存储转发的包交换技术和分布式路由策略，支持 BE 服务。Teraflops 是 Intel 面向通用计算领域研制的众核处理器，包含 80 个处理器内核，采用 8×10 的 2D 网格拓扑结构，网络采用虫孔交换技术和基于轮转调度的通道仲裁策略，链路采用 5 级流水设计。

除了上述传统的 2D NoC，3D NoC 逐渐成为 NoC 研究中的一个重要分支。将 3D 集成电路 (Three-Dimension Integrated Circuit) 技术应用到中，产生了 3D NoC 技术。这种技术将多个晶片在垂直方向进行堆叠，层间通过高速且高密度的硅通孔 (Through Silicon Via, TSV) 相连，其基本构成主要包括路由器、通信链路和资源节点 (处理器单元或 IP 核) 等。

3.2.4 SIP 技术

SIP (System in Package) 是近几年来为适应系统小型化以及模块化需求而出现的集成技术。SIP 利用已有的电子封装和组装工艺，组合多种集成电路芯片与无源器件，封闭模块内部细节，降低系统开发难度，具有体积小、成本低、开发周期短、系统性能优良等特点，目前在工业控制中得到广泛应用。当前，随着集成化程度的提高，电路面积持续减小，同时芯片的对外引脚持续增加，现有的二维封装结构面临着挑战，SIP 采用三维封装技术，能够有效减小电路面积，增加对外引脚数量。SIP 封装并无一定形态，就芯片的排列方式 SIP 可为多芯片模块的平面式二维封装，也可利用三维封装结构，以有效缩减封装面积。SIP 内部接合技术可以是单纯的打线接合 (Wire Bonding)，亦可使用覆晶接

合（Flip Chip），但也可二者混用。除了二维与三维的封装结构外，另一种以多功能性基板整合组件的方式，也可纳入 SIP 的涵盖范围。此技术主要用于将不同组件内藏于多功能基板中，以达到功能集成目的。与印制电路板上的电路设计相比，SIP 能最大限度地优化系统性能、避免重复封装、缩短开发周期、降低成本、提高集成度；与 SoC 对比，SIP 具有灵活度高、集成度高、设计周期短、开发成本低、容易进入等特点。

早在 20 世纪 80 年代，美国在无线通信方面率先开展三维微波集成电路系统级封装技术研究，自 1987 年开始实施了微波毫米波单片集成电路（MIMIC）计划。欧洲的法国和英国等国家自 20 世纪 90 年代中期以来，也积极开展三维集成电路与系统技术的研究，特别是在航空、航天电子领域，取得了一系列研究成果。

最值得一提的是，法国的 3D-plus 公司采用筛选和叠层封装技术，走出了一条研发宇航级存储器件的成功之路。3D-plus 采用 SIP 技术，专利技术遵循从标准封装到裸片和晶圆级这样一种发展进程，产品几乎覆盖了宇航级存储器器件，包括 PROM、E2PROM、Flash NAND、Flash NOR、MRAM、SDRAM、DDR1 RAM 和 DDR2 RAM 等，客户（包括伽利略计划在内）遍及世界各地，在我国的航空、航天领域也得到了大量应用。

3.2.5 嵌入式操作系统

在桌面应用领域，Windows、Linux 以及 UNIX 基本囊括了绝大部分的份额，而在嵌入式实时操作系统领域，其种类超过百种，目前在市场上占主流的有 VxWorks、Integrity、QNX、RTEMS、ThreadX、LynxOS、RT-Linux、uc/OS，这些嵌入式实时操作系统都有着广泛的应用。

目前，针对单核处理器的嵌入式实时操作系统已经较为成熟，以 VxWorks 为例，它支持多种处理器体系结构，具有高效的可裁剪内核，硬件驱动较为完善，配套 IDE 软件功能强大，支持图形化的应用开发、模拟器调试以及目标机调试。针对逐渐兴起的多核处理系统，有越来越多的嵌入式实时操作系统加入到多核的阵营。但是，目前支持多核处理器的嵌入式实时操作系统还为数不多，主要有 WindRiver 公司的 VxWorks，QNX 软件系统有限公司的 Neutrino，用于美国国防系统的 RTEMS（Real Time Executive for Multiprocessor Systems），由 Linux 发展而来的 MontaVista Linux，以及由 Red Hat 公司开发的开源 eCos（Embedded Configurable Operating System）等。此外，对有些简单的单核操作系统进行改造也能够对多核处理器进行支持，例如对 μC/OS 进行改造，增加简单的中断和共享内存通信机制，使它可以运行于非对称多处理器环境。

VxWorks 是美国 WindRiver 公司开发的实时多任务嵌入式操作系统，是目前应用最为广泛的一个嵌入式操作系统^[27]。VxWorks 通过处理器可选库 VxMP 最多可支持 20 种处理器，主要有 680x0、683xx 系列、SPARC、SPARClite、Intel 80960 系列、MipsR3000 和 R4000 等。VxWorks 对多核处理器的支持最初采用非对称多处理（AMP）方式，允许不同处理器上的任务进行同步和数据交换；后续版本（例如 V6.7 和 V6.8）增强了对对称多处理（SMP）的支持，并扩展了开发工具，能够在多核处理器上进行 SMP 和 AMP 系统的调试和开发。

QNX 是一个真正的微内核操作系统，它的驱动程序、应用程序、协议栈和文件系统都在内核外作为进程运行，使得几乎所有组件都可以失败再重启，而不影响内核或其他组件^[28]。QNX 使用消息传递作为进程间通信（Inter Process Communication, IPC）的基本方式，是目前唯一一个全面支持多核解决方案的嵌入式操作系统，它支持 SMP、AMP 和 Bound 多处理等模型。它还提供了一套用于开发多核程序的工具，包括监控核间通信工具，应用 profile 工具，多个核上的源码级调试等。

RTEMS 是一个开源的实时嵌入式操作系统，最早用于美国国防系统，早期的名称为实时导弹系统（Real Time Executive for Missile Systems），后来改名为实时军用系统（Real Time Executive for Military Systems），现在由 OAR 公司负责版本的升级与维护^[29]。对于多处理器系统的支持，RTEMS 提供了一系列灵活有效的机制，支持的多处理器系统不但包括传统的紧耦合系统与松耦合系统，还对混合耦合系统、异构系统以及混合异构系统都提供了强有力的支持。RTMES 屏蔽了底层通信的细节，允许整个多处理器系统的软硬件在逻辑上表现为一个系统。目前可以利用其原有的 AMP 模式支持多核，然而，这种模式需要开发人员静态地分配任务到每个处理器内核上，这不仅增加了开发人员的负担，而且没有充分发挥多核处理器的性能，当不同处理器内核上任务间的通信过于频繁时，系统性能就会明显下降。

eCos 是由 Red Hat 公司开发的一个开放源码的可配置、可移植、嵌入式实时操作系统。eCos 支持的主要处理器有 Motorola 6800、PowerPC、MIPS、X86、ARM 和 SPARC 等。其对多处理器的支持以 SMP 的方式实现，所有处理器共享一个存储器空间，处理器之间的通信通过共享变量或者消息传递完成，任务采用统一调度的方式实现。但其对 SMP 的支持具有一定的局限性，需要在选定的处理器构架和平台上提供支持，且对硬件有一定的限制。例如针对 SPARC 体系结构的处理器，eCos 最多支持的内核为 8 个，典型值为 2 个或 4 个，且要求硬件提供比较和交换指令的同步机制，中断控制器需将中断信号送给每个内核，硬件提供缓存一致性维护，共享内存且对所有内核地址一致等。

除了上述较为通用的嵌入式操作系统都对多核进行支持以外，一些专用的和研究领域的嵌入式操作系统也针对多核处理器开发了相应的版本。例如 Karlsruhe 科技大学设计了基于 SMP 架构的 L4 微内核操作系统^[30]，提出了多处理器兼容用户态调度器，通过调度器可把某个线程从某个处理器转移到另一个指定的处理器上执行。

随着多核和多处理器系统的发展，为了实现整个系统的高可靠运行，提供对处理器间或者处理器内核间容错的正常，在处理器间的任务迁移技术逐渐成为嵌入式实时操作系统研究的一个新热点。任务迁移一般分静态任务迁移和动态任务迁移，前者一般指常规的任务调度，绑定在某个处理器内核或处理器上，后者则基于负载、功耗、容错的需要，把任务动态迁移至别的处理器或节点上，比较复杂。关于任务迁移的研究最早出现在分布式计算机领域，Robinson 等人详细介绍了并行计算机系统中基于消息传递接口的一种任务迁移机制。他们在消息传递接口中添加了称作 Hector 的任务动态分配器。Hector 的设计基于一种主从层次结构。主节点上称为主分配器的任务负责任务迁移决策，同时与从分配器进行通信。各个从节点上都有一个从分配器监控 MPI 任务的运行性能特

征，并把信息反馈给主分配器。在分布式计算机中迁移任务时需要传输任务的状态以及任务迁移的过程，迁移的过程主要包括打包任务状态、传输任务状态以及任务的恢复等步骤。另一种任务迁移基于 MPSoC 或者 NoC，其通信资源以及各个节点的存储资源都非常有限，而在任务迁移过程中，整个任务传输有较大的通信开销，因此在分布式计算机中的任务迁移研究结果并不能直接移植到 MPSoC 或者 NoC 中。为了能够在多核系统中合理地评估任务迁移的开销，需要研究和实现新的任务迁移算法和迁移机制。目前，该项技术处于实验室的研究阶段，还没有成熟的商用嵌入式操作系统具备该项功能。

3.3 国内研究进展

3.3.1 片上系统

随着集成电路与 IP 技术的快速发展和日臻成熟，国外采用 SoC 技术进行设计已经成为一种主流技术。国内也开始采用 SoC 进行电路设计，目前发展趋势非常迅速。当前国内各个有实力的企业、高校和科研院所都在研发自己的 SoC，但往往带有一定的领域专用特征，或者是在国外知识产权（IP）的基础上进行开发。

（1）商用 SoC

华为在通信领域研制开发了多款 SoC。例如海思 K3V2，该颗芯片是全球首款四核 ARM SoC，其正式发布时间比 NVIDIA 的 Tegra 3 早了一天，芯片采用 40nm 工艺，GPU 模块使用了体积最小、效率最高但是非常少见的 Vivante GC4 000，这使得 K3V2 拿下了业界最小四核处理器的殊荣，在华为当时的高端手机上大量使用。但这款芯片发热量大并且频率不易提升，兼容性不佳。随后，研制出基于 28nm HPM 工艺的 Cortex-A9 四核 SoC，以及四核 A15 + 四核 A7 的（高端八核）SoC，除了与 ARM 内核类似的通用处理和运算功能之外，两者均支持 GSM/WCDMA/TD-SCDMA/TD-LTE/FDD-LTE 等多种制式的移动通信网络。在 2014 年华为了改变在 SoC 方面的劣势，布局了全新的产品，即代号为“Kirin”的全新麒麟系列处理器。以定位主流的麒麟 910 系列为例，芯片采用 28nm 工艺，具有漏电率低，频率表现出色等特点。在 CPU 模块方面，麒麟 910 采用了 ARM Cortex-A9r4 的四核配置，并具有针对存储设备的管理以及增强的指令跳转预测。在功耗控制方面，也提供一些更为精确有效的节能功能，最高频率可以达到 2.3GHz。GPU 采用 Mali-450MP4 GPU，支持 OpenGL ES2.0。在内存配置方面，支持单通道内存控制器，最高可支持 LPDDR3 1 600，内存带宽为 6.4GB/s。在网络方面，支持目前几乎所有制式的 2G、3G、4G 网络，尤其是支持 LTE Cat4 150Mb/s 4G 网。

中兴公司的 SoC 研发也面向移动终端和通信领域。该公司研发 WiseFone 7510 LTE 基带芯片首款 4G 基带处理芯片，采用 28nm 工艺，在芯片网络制式方面支持 TD-LTE/LTE FDD/TD-SCDMA/GSM 商用芯片，支持 R9 LTE Cat 4 高速传输。其升级版（WiseFone 7550s）则是一款八核的 SoC，采用 ARM Cortex-A7 八核架构，八核 CPU 部分在高频率运行的同时保持低功耗，与当前主流四核英伟达 Tegra 4、高通骁龙 800 相比，功耗降低了 60%。

(2) 宇航级 SoC

在宇航级 SoC 的研究领域,以航天科技集团公司下属的多个院所为代表的单位在 21 世纪初就开展了相关工作。2005 年前后开始研发航天 SoC 产品,主要基于开源的 SPARC V8 架构,研发了多款 SoC,开始应用于武器装备和航天器。北京控制工程研究所自主研发了适用于下一代卫星的小型 SoC 控制器。控制器的核心芯片 SoC2008 把原来整数处理单元、浮点数处理单元、存储器管理器和 1553B 等电路功能集成到芯片内部,芯片外形如图 1 所示^[31]。SoC2008 芯片的性能与 Aeroflex Gaisler 2010 年推出的容错版芯片 UT699 性能相当,达到国际先进水平。目前,SoC2008 已经用于某试验卫星中心管理单元中,自 2012 年作为控制计算机主体发射升空以来迄今为止工作正常。目前,SoC2008 已经应用于多颗导航卫星的敏感器处理单元,以及多颗微小卫星星座的综合电子单元等产品。

2010 年起,北京控制工程研究所在 SoC2008 的研究基础上,开始研制集成四个 SPARC V8 内核的 SoC2012 芯片,如图 2 所示。SoC2012 项目得到了导航卫星系统关键技术攻关项目的支持,在 2012 年年底完成流片,在基于四核并行容错 SoC2012 的导航卫星中进行搭载飞行试验,预计卫星于 2015 年左右发射。



图 1 集成 SPARC V8 内核的 SoC2008 芯片

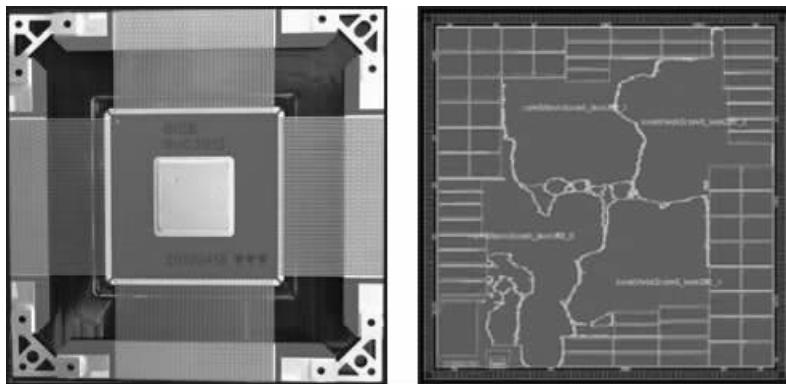


图 2 集成四个 SPARC V8 内核的 SoC2012 芯片及版图

3.3.2 嵌入式操作系统

在嵌入式操作系统方面,国内的研制单位也集中在大型企业以及科研单位。

北京控制工程研究所面向航天应用开发了具有自主知识产权的空间高可靠嵌入式实时操作系统,从 21 世纪开始,已经研发出多个版本,例如 SpaceOS I、SpaceOS II 和 SpaceOS II 多核版。其中 SpaceOS I 已应用到我国在轨的数十颗飞行器,它们全部在轨稳定运行。该版本操作系统支持多任务调度,支持中断的多级响应机制,支持静态内存分配,支持星船软件在轨编程,以及看门狗管理和超时处理等容错功能。

随着我国空间任务多样化、复杂化，像月面巡视器、火星探测器、空间站等重大复杂空间应用对软件提出了更高要求。北京控制工程研究所开发了 SpaceOS II，该操作系统具有以下特点。

- 支持性能优越的动态内存管理方法，分配和回收时间短，并且碎片较小；
- 基于多优先级队列的时间片轮转调度方法，支持同优先级任务运行，能够提供更多的任务数量；
- 实现快速中断响应方式，适用于发生频率较高而且中断服务程序处理时间较短的中断；
- 信号量和消息队列实现了任务间的同步和通信机制，能够满足多个任务对相同资源和外设的互斥操作及任务间的时序控制；
- 程序代码经过了大量的优化，并采用了剪裁技术，内核可精简；
- 良好的可扩展性：系统的功能可以通过创建任务来方便地扩展，系统软件代码修改和升级的代价小。

SpaceOS II 已经成功应用于我国嫦娥三号月面巡视器中，圆满完成了月面巡视任务。

在 SpaceOS II 的基础上，北京控制工程研究所又成功研制了支持多核片上系统的多核操作系统。该操作系统面向多核环境，支持多核 SoC 的 BSP、支持多核任务并行调度，支持多核任务间通信。目前它已经应用于导航卫星的计算机产品中，计划将于 2015 年发射。

此外，中国电子科技集团公司第三十二研究所自主研制了第一个提供 VxWorks 兼容接口的嵌入式实时操作系统 ReWorks，该系统采用面向对象和微内核技术开发，特别提供了 VxWorks 兼容层，具有可裁剪、实时性、安全性和稳定性好等特点^[32]。

北京银科京城公司开发的 Delta OS 是一个具有完全自主知识产权的嵌入式实时操作系统，它具有内核精简、功能模块可裁剪、系统参数可配置、实时性强、可靠性高的特点，可以广泛应用于工业控制、国防，以及民用的信息家电和个人数字终端等消费类产品^[32]。

2012 年 3 月，北京计算机技术及应用研究所成功研制出基于“龙”芯处理器的移动操作系统，即第二代“天熠”移动操作系统，实现了航天领域首款国产化移动软件产品。该系统拥有自主知识产权，提供了良好的图形界面和用户体验，可操作性强，具有强大的网络功能，可实现移动通信中数据的快速准确交互。

3.4 国内外研究进展比较

由国内外的研究进展容易看出，近年来嵌入式系统发展非常迅速，国内企业和研究机构的实力得到了大幅度的提升，与国外同行的差距正在逐渐缩小，但还存在较为明显的差距。

在 SoC 技术方面，国内 SoC 技术多以国外开源的 IP 或者向国外购买的 IP 为基础进行开发，真正具有完全自主知识产权的 SoC 很少，例如国内通信界的华为 SoC 也是基于

ARM 体系结构进行开发的。很多应用都是较低层次的，还没有达到系统级的水平。

在嵌入式操作系统方面，国内已经具备了多个具有自主知识产权的操作系统，满足了行业内的实际需要。但是存在编译和调试环境不配套，系统研发和应用各自为战，推广不够的问题。在这种情况下，软件的正确性和可靠性验证都较为局限。在多核并行操作系统开发方面，国内还处于起步阶段，与国外成熟的操作系统还有差距。

3.5 发展趋势与展望

当前，工业控制领域嵌入式系统的发展方兴未艾，随着工业控制领域不断出现的新需求，嵌入式系统将朝着通用化、模块化、系统化、并行化、网络化、智能化、更高集成度和更低功耗的方向发展。

在芯片集成度方面，向着集成更多功能、更低电压、更低功耗的 SoC 方向发展，而且不同行业都逐渐形成自身的多功能集成的 SoC。例如，Intel 的夸克针对与人们生活密切相关的智能设备，航天领域则开发并应用了抗辐射且集成了航天专用 Spacewire 总线的 SoC。此外，芯片集成由两维扩展为三维，业界不断出现并成功应用 SIP 技术。

多核处理器由最初的双核、4 核朝着单芯片多核、众核方向发展，甚至发展为片上网络。目前已有多达 64 核或者更多核的处理器。而且目前多核处理与 SoC 相融合，多核片上系统已经在多个领域研制成功，并得到实际应用。

高度集成 SoC 和多核处理器的快速发展为操作系统的发展提出了挑战，片上多核和片上网络为操作系统的发展又开发了一片广阔的天地。这需要把以前并行操作系统、分布式操作系统和网络操作系统的概念缩聚到一个芯片上，同时需要保证一定的功能和性能，保证任务调度的均衡性。这都给操作系统的研究和开发带来了新的课题。

4 高可靠容错计算机关键技术

4.1 概述

高可靠容错计算机是工业控制计算机的一个特殊领域，这类工业控制计算机由于其运行环境的苛刻性，比普通工业控制计算机具有更高的可靠性和安全性，往往需要进行特别的容错设计。这类特殊的工业领域一般是无人值守或者危及安全的应用领域，例如核工业控制、航空和航天等，一旦计算机发生故障，将会带来巨大的损失甚至灾难。因此，人们希望自己所使用的计算机系统是个高可靠的系统，即使发生了故障也能够正常工作或者基本正常工作，至少不产生严重后果。高可靠、高安全的容错计算机技术正是在这种需求的牵引下产生和发展起来的。

为了实现计算机的高可靠性和高安全性，产生了多个研究领域和学科分支。目前，在金融、通信、核工业、军事工业、航空和航天等高可靠工业控制领域都采用了容错计算机及其相关技术，以保证系统的可靠性和安全性。在容错计算机领域，最基本的原理就是冯·诺依曼提出的用低可靠性器件以冗余方式构造高可靠性的系统，在计算机内部出现故障的情况下，整个计算机系统仍能正确工作。在很多经典容错系统中，都采用了多模冗余技术来实现系统的可靠性设计，例如美国 NASA 航天飞机中的五模冗余控制计算机。可编程器件出现以后，使得硬件可以在线改变结构和功能，这种可重构技术迅速成为研究的热点，而采用可重构技术实现的硬件容错技术目前已经成为容错领域一个重要的研究方向。特别是，近年来各类仿生学不断与计算机技术相结合，甚至产生了基于智能方法的容错技术。在这类高可靠、高安全的工业控制计算机中，软件的可靠性和可信性也是一个重要的研究方向。以往软件的可靠性和可信性一般采用多级测试体系来保障。目前，随着工控领域软件规模急剧扩张，原有测试方法工作量剧增，而且其测试结果的可信性越来越不能满足需求，形式化方法逐渐成为高可靠、高可信软件领域中一个颇具前景的研究方向。

4.2 国际研究现状

4.2.1 高可靠计算机体系结构

第一代计算机（1946~1957 年）由电子管、继电器和延迟线存储器等构成。由于这些元件的失效率相当高，且容易受瞬时故障的影响，造成系统的平均无故障时间很短，只有几分钟到两三个小时，为此需要采用故障检测和恢复技术。例如该时期的 IBM 650、UNIVAC 等计算机均采用奇偶校验等措施来提高可靠性。容错技术的最典型应用是航空航天领域，早在 20 世纪 60 年代这种技术已经开始用于航天领域。高可靠容错计算机的经典应用有 1969 年美国喷气与推进实验室（JPL）为阿波罗（Apollo）计划研制的容错计算机 STAR（Self-Testing And Repairing），该计算机采用故障检测模块来发现故障，并通过替换永久故障单元来容忍永久故障，这是早期容错计算机研发历史中的里程碑^[34]。

在目前高可靠容错计算机中，容错结构包括模块冗余和系统冗余。从硬件容错来讲，按照备份方式分为热备份、冷备份；按照冗余的数量分为双机备份、多机备份；按照故障恢复方式分为静态冗余和动态冗余。选取何种冗余方式，主要根据系统的实际要求来确定。对于故障处理实时性、可靠性和安全性要求高的系统，一般采用静态冗余方式，典型的如三取二表决方式；对于寿命要求高的系统，可采用冷备份方式；对于可靠性要求特别高的系统，可采用混合冗余的方式，即静态和动态相结合的方式^[35]。

（1）模块级冗余结构

采用模块级冗余结构的系统是容错系统结构中出现最早的系统。计算机系统一般

可以分为 CPU、存储器和 I/O 模块，模块级冗余结构可以通过在这些模块上的冗余来实现故障的检测和处理。模块级冗余的形式可以有双重备份、多重备份的形式。相对于同样结构的系统级冗余方式，这种系统一个显著的优点是可靠性比较高。但是由于模块之间的重构比较复杂，故障处理难度大，系统中共用部分较多，存在的单点失效模式比较多。早在 20 世纪 60 年代，由于电路的集成度不高，世界上第一台航天用容错计算机（即 STAR 计算机）就采用模块式结构的容错系统。STAR 计算机曾应用在阿波罗飞船上，这种结构的系统是通过对整个计算机模块的冗余来实现系统可靠性的提高^[34]。

（2）备份冗余结构

备份冗余结构是常见的容错系统结构之一，相对于模块级容错计算机系统，其主要区别是以单个计算机系统作为备份单元进行备份。典型的方式有冷备份方式和热备份方式，以及温备份方式。温备份方式可以视为热备份方式的一种变形，在实际系统中这种方式比较少见。

1) 冷备份容错结构是指系统中只有一个机器加电工作，其余机器不加电的系统。当工作机发生故障时，则通过容错切换到备份机，并使其加电到工作状态。

2) 热备份容错结构可以提高检测故障的实时性，在这种结构中，两机相互比较来检测故障。与冷备份结构的主要不同是，备份机加电且相互之间存在通路，以实现关键数据的相互交换和对比，双机并行运行相同的软件，两机输入端并联，输出端由双机切换控制电路连通，只允许其中一个计算机的输出信号送往外部。

20 世纪六七十年代，美国典型的容错计算机 JPL 旅行者（Voyager）深空探测卫星容错计算机，采用双机热备份动态冗余容错机制，在太空运行了 35 年。在关键领域的容错服务器，例如金融、电信、民航、交通、电力、制造等行业，广泛采用 HP 和 IBM 的双机热备份冗余结构，对应用层透明，但这种服务器价格较高，且切换时间具有一定的不确定性。

（3）多模冗余结构

在实时性要求比较高的控制系统中，需要采用静态冗余和表决的方式来提高可靠性，即多模冗余结构。采用这种方式的典型代表是三机冗余 TMR 结构系统，该结构采用三取二策略实现系统容错，系统中只要有两个机器正常工作时就可以输出正确的结果。硬件表决器具有直观、快速的特点，但同时它本身的表决器是一个单点。而且，随着输出路数增多，表决器越来越复杂，可靠性随之降低。美国 NASA 的航天飞机容错计算机采用五模冗余的结构，在航天飞机二十多年的运行期间，该容错计算机成功地完成了飞行任务^[36]。其他成功的应用还有采用三片 FPGA 实现三模冗余容错系统，并能够在一模发生故障的情况下进行离线故障检测^[37]。

（4）其他容错结构

除了上述几种容错结构外，还有冷备份与热备份结合的混合容错系统结构，以及由软件实现的容错系统结构，即软件实现的容错（Software Implemented Fault Tolerance，SIFT）。

4.2.2 可重构技术

可重构技术是 1963 年由美国加利福尼亚大学洛杉矶分校的 Gerald Estrin 等研究者首先提出的。经过几十年的发展，现代计算机体系结构从本质上讲并没有太大的变化，仍然以冯·诺依曼结构为基础。而可重构计算（Reconfigurable Computing）技术结合了 ASIC 芯片和 CPU 两者的优势，既具备接近于 ASIC 芯片的速度和效率，又具备类似于 CPU 的通用性和可编程特性。可重构系统的总体思想是以时间上的代价，来实现空间资源重用的目的。

目前在工业控制领域中 FPGA 大量使用，这也为可重构技术的应用提供了便利。目前，可重构容错技术主要基于 SRAM 型的 FPGA，其中具有代表性的是 Xilinx 公司的 Virtex 系列 FPGA。该公司提供了全局重构、动态刷新、动态部分重构等多种手段。可重构容错技术可充分利用 FPGA 这些特点，大幅度提高特殊工业控制领域中计算机的可靠性和安全性。

除了在 FPGA 片内实现重构之外，研究人员也提出了多种可重构体系结构。加州大学伯克利分校研究人员提出了一种混合结构的可重构系统 GARP，其目标是将可重构计算单元与普通 RISC 处理器集成到一块芯片上，借以探索可重构计算对于应用的加速能力^[38]。GARP 采用了 MIPS 作为其主处理器的原型，其可重构部分采用 Xilinx 4000 系列构成阵列，同时规定每行必须包括 24 个模块。美国麻省理工学院在 1997 年提出了一种网格型计算阵列结构——RAW (Reconfigurable Architecture Workstation) 结构，主要目的是为了设计一种简单的、高度并行的 VLSI 结构，使得编译器能够充分评估、优化和利用硬件资源^[39]。卡内基 - 梅隆大学提出了 PipeRenew，在设计时引入了一项很有特色的配置技术，即流水线重构，与 CPU 设计中的流水技术类似，其思路是将一个完整的配置过程拆分成不同阶段，当后续阶段在配置时，前面已配置的部分已经可以运行了^[40-43]。

进入 20 世纪 90 年代后，出现了把仿生学思想运用到重构中的多种智能重构技术，这种技术代表着高可靠容错计算机的一个未来发展方向。智能重构技术能够模仿生物界的机制实现硬件重构，从而实现硬件自主检测故障、自主修复故障，甚至可以根据环境自适应外界环境的变化。在智能重构技术中，最具代表性的有可进化硬件（Evolvable Hardware, EHW）技术和人工免疫硬件（Artificial Immune Hardware）技术。可进化硬件的概念产生于 20 世纪 90 年代，最初是 1992 年由日本的 Hugo de Garis 和瑞士联邦工学院的科学家同时提出的。最初的研究主要集中在欧洲和日本^[44]，但是美国由于 NASA 和 DoD（国防部）的支持发展十分迅速，现已处于国际领先地位，JPL 甚至开发出专用的现场可编程晶体管阵列（Field Programmable Transistor Array, FPTA）用于该方向的研究^[45]。人工免疫系统是一种受生物体免疫系统工作流程和机理启发而发展起来的一种人工智能系统，具备生物体免疫系统异体检测与识别、异体清除和系统修复能力。具有高可靠需求的计算机可以借鉴这种技术实现故障修复和容错。1994 年，美国新墨西哥州立

大学的 S. Forrest 教授提出了用阴性选择算法 (Negative Selection Algorithm, NAS) 来模拟生物体免疫系统异体识别功能，以达到检测计算机文件是否被病毒所篡改的目的^[46]。在航空领域，Perhinschi 等人针对飞行器控制系统构成复杂、故障模式多样等特点，提出了利用人工免疫系统进行飞行器控制系统故障检测定位与评价 (Failure Detection, Identification, and Evaluation, FDIE) 方法^[47]。

4.2.3 可信软件

在工业控制领域，提高软件可靠性、构建高可信软件系统已成为世界范围的重要课题。一般认为：软件的可信 (dependable) 是指软件系统的运行行为及其结果总是符合人们的预期，在受到干扰（如操作错误、环境影响和外部攻击等）时仍能提供连续的服务。

早在 1990 年，美国 NASA 举办了第一届形式化方法研讨会 (NASA Formal Methods Symposium, NFM)，鼓励学术界和工业界开发形式化方法用于验证航天领域中关键安全软件的正确性，从而为研发可靠的自动机器人、下一代航空交通工具和航天飞船打下基础。在其他的一些危及安全的工业控制领域中，欧洲很多国家早就开始使用形式化方法提高软件系统的可信性，例如法国城际铁路的 SACEM 系统^[48]和空客公司的航空电子设备软件^[49]等。

形式化验证是一种严格保证软件系统可靠性的重要技术。通过逻辑推理的方式来证明软件正确性的想法，早在 20 世纪 60 年代就由 Floyd、Hoare 等人提出。近年来，由于程序设计语言理论和验证理论研究的突破，自动定理证明和程序分析能力提高，形式化程序验证又重新成为当前研究的热点。

20 世纪 90 年代后期，英、美、法、德等国相继出现了提供形式化规约和验证技术产品的高科技公司，著名计算机公司如 Intel、IBM、Microsoft、Lucent 纷纷在其产品开发中使用形式化方法。美国政府在利用形式化方法研究高可信软件系统方面给予了大力支持：美国 NASA 于 1995 年 7 月和 1997 年 5 月先后发布《形式化方法规范和验证指南》，用于指导航天软件的开发；美国国防高等研究规划局 (DARPA) 资助的 Formal Methods 计划，重点支持对高可信软件系统的原理和相应支撑工具的探索；美国国家自然科学基金会在 2000 年财政年度启动了至今为止规模最大的研究计划 Information Technology Research，强调利用验证和形式化方法来保证系统行为的正确可靠。在欧洲，形式化方法自 2008 年起被正式列入欧洲航天局航天数据、控制及软件系统研讨会 (ESA Workshop on Avionics Data, Control and Software Systems, ADCSS) 的讨论内容，此后每年在此论题上都会有若干与形式化验证方法相关的论文发表。2008 年，欧洲航天局 (ESA) 投入近百万欧元与欧洲数个大学和研究机构合作研究能够确实保障软件可靠性和安全性的软件开发方法，该项目被命名为 COMPASS (Correctness, Modeling and Performance of Aerospace Systems)，经过两年的开发，该项目有了初步的成果并被应用于卫星平台的开发上^[50]。至 2011 年，欧洲航空研究技术中心正式将把形式化验证方法列入软件开发的标准流程提上议程。此外，欧洲航天局不断尝试形式化方法的应用：火星探测任务 (ExoMars) 中使用的机器人监控系统中引入了形式化方法；由瑞典研发的星球漫步者 (MarsRover) 的设计验证方案

中使用形式化验证技术将系统控制器的基础部分使用 C 代码实现。越来越多的实践证明，形式化方法在系统设计和验证，特别是高可靠性系统的设计和验证中，有着不可取代的地位。

4.3 国内研究进展

4.3.1 高可靠计算机体系结构

国内对高可靠容错计算机的研究多集中在高校以及具有特殊应用领域的科研机构。哈尔滨工业大学、清华大学、北京控制工程研究所和中国科学院计算技术研究所等机构较早开展高可靠容错计算机的相关研究，出版了高可靠计算机方面的多部专著，例如杨孝宗著的《容错技术与 STRATUS 容错计算机》，每年发表多篇论文。在工程应用方面，国内航空航天领域走在了高可靠容错计算机研究的前列。北京控制工程研究所从 20 世纪五六十年代开始研制高可靠容错计算机，并成功应用于多颗卫星和飞船上，容错方式包括模块级冗余、双机冷热备份以及三机热备份。其中，最值得一提的是，在神舟飞船中采用了改进型的三机热备份结构，保证了飞行任务的圆满成功。在民用高可靠服务器领域，目前美国的高端产品都在中国生产和加工，例如 IBM、HP 等，但具有我国自主知识产权的容错服务器研制还很少，以联想集团为代表的高可靠产品所占的市场份额还很低。

4.3.2 可重构技术

可重构技术方面相对国外起步虽然较晚，但发展十分迅速，目前各个高校和科研院所都有相关研究团队。由于 FPGA 这种器件的易用性，国内基于该器件的相关研究非常多。在对可靠性要求很高的工业控制领域，在 FPGA 基础上进行可重构容错是研究的重要方向。目前，航天系统的科研单位、中国科学院等都进行了较为深入的研究，尤其是针对 Xilinx 公司的 Virtex 系列 FPGA，并在工程上应用了动态刷新、全局重构和部分重构等多种技术。20 世纪末，国内可重构技术研究的重要对象是采用 Xilinx 公司的 XC6000 系列 FPGA。21 世纪初，发展为 Virtex、Virtex II 系列 FPGA，并成功研制基于这几个系列器件的全局重构技术和动态刷新技术，有效解决了容错以及多配置 FPGA 的问题。近年来，Xilinx 公司加大对可重构技术的支持，把部分重构的 EDA 工具软件 Plan Ahead 集成到 ISE 中，使得 Virtex 4/5/6 等后续系列 FPGA 能够更好地支持部分动态重构。

北京控制工程研究所针对空间高可靠控制计算机，在国家级的预研课题中进行可重构研究，并在航天型号任务中进行实际工程应用。该单位研究和实现了多种重构，包括基于 Xilinx 公司 Virtex、Virtex II、Virtex 4 和 Virtex 5 系列 FPGA 的全局重构技术、工作时动态刷新技术、动态部分重构技术，并把上述技术应用于多个型号的航天器的高可靠计算机容错设计中，使得计算机产品取得了圆满成功。

在智能重构方面，国内的研究机构主要包括西安电子科技大学、中国科技大学、南京航空航天大学、武汉大学、中国地质大学、北京控制工程研究所等单位。研究对象从

最初的门级电路重构发展到模块级电路的重构，并不断尝试把这种技术应用到实际的工程中。

4.3.3 可信软件

国内在可信软件的研究中，北京控制工程研究所针对航天软件的可靠性做了大量工作。该单位得到了国家自然科学基金支持，面向航天软件，特别是操作系统，进行了可信研究和验证工作。针对操作系统自身的特点，该单位提出了异构的验证方法和技术，克服了系统的复杂性，提出了一个支持可信航天器操作系统软件构造和验证的开放性框架。该框架支持采用异构技术来验证系统中的不同模块，并将各模块统一组装成可信软件系统。框架具有通用性和开放性。它将从四个方面为可信航天器操作系统软件的构造和验证提供支持，即：①系统结构和抽象层次的形式化描述；②具体抽象层次的领域专用语言和逻辑的开发；③各类工具的开发、验证和集成；④支持不同验证方法集成的基础语义和逻辑框架。在软件开发过程中，按照需求、设计和实现阶段，按照自底向上的层次进行验证，并采用 Event-B 形式化方法对系统进行了建模和验证。

中国科学院软件所在嵌入式操作系统的可靠性验证方面进行了系统化的研究工作，唐稚松教授开发了时序逻辑语言 XYZ 以及相关 CASE 工具，并对某嵌入式系统进行了验证分析，发现其中存在的优先级翻转问题。中科院软件所国家基础软件研究中心针对嵌入式操作系统的验证技术包括：基于抽象解释技术对驱动模块的验证方法，基于动态二进制代码转换技术和符号执行技术对单元模块的系统测试等。林惠民教授在进程代数、模态逻辑和 μ 演算等方面进行了大量的研究。张文辉研究员针对控制模型检验中的内存消耗峰值进行了研究。

国防科技大学对 System V 进程通信机制进行了形式化验证，也在量子密码协议安全性验证研究上使用了形式化分析方法。华东师范大学可信嵌入式软件分析与验证实验室的蒲戈光教授等人研究了基于符号执行的应用于单元测试的测试用例自动化生成技术以及基于源代码和二进制码操作系统的验证技术。

南京大学在需求级软件形式化与自动化技术方面进行了研究，提出了基于 UML 的需求模型验证方法；把面向对象的建模语言 UML 与形式化方法 B 相结合，将 UML 类图转换为 B 方法的抽象机，降低了形式化方法的使用门槛。此外，它还设计了结构化需求定义语言 NDRDL 和对象式需求定义语言 NDORDL，研制了相应的需求级自动化系统。南京大学钱振江博士研究了安全操作系统的形式化设计与验证方法，提出了操作系统对象语义模型（OSOSM），采用分层结构将操作系统中的行为主体和资源抽象为操作系统对象，建立操作系统的论域，利用以操作系统对象变元集合为定义域到论域的映射表示操作系统的状态，描述操作系统系统调用等行为的语义。使用逻辑系统的谓词公式表达操作系统的安全属性，给出如何验证操作系统在运行过程中保持安全策略和属性的形式化描述方法。

中国科技大学陈意云教授研究了程序检验的方法在内存地址安全、多线程数据竞争等方面的应用。郭宇教授对操作系统内核验证方面做了相关研究，包括对内核雏形结构

的验证方法与线程切换，可以处理某一类的特殊内核结构，针对实际内核，提出了一种程序逻辑系统，可以验证基于线程切换的并发内核代码，但尚不支持线程管理、中断管理、多核等其他实际内核特性。

4.4 国内外研究进展比较

综合国内外研究现状，可以看出我国在高可靠容错计算机技术方面与国外也存在一定差距。在高可靠容错计算机体系结构方面，我国各个行业的发展路线不尽相同。在航天等高可靠、高安全领域能够满足目前的工程实际的需要，与国外基本处于同一水平；但是在民用高可靠计算机方面与国外差距还很大。

在可重构技术方面，国内研究很大程度上基于国外 Xilinx 公司的 FPGA 产品，由于该产品的技术细节不为国内所掌握，因此研究具有较大的困难，虽然取得了很多技术上的突破，但研究的进展和成果很大程度上有赖于厂家资料公开的进程。这种情况下，也不能形成具有我国自主知识产权的技术。自 1986 年 Xilinx 公司发明可编程器件 FPGA 以来，已有近 30 年的时间，研制具有中国完全知识产权的大规模可编程器件已经迫在眉睫。

在可信软件研究领域，我国与国外也有相当大的差距，目前一些形式化验证通用标准以及流行工具大多由国外提出，我国在这些方面还比较欠缺。值得高兴的是，随着软件可靠性的迫切需求，我国近年来形式化方法发展比较迅速，在模型的提出、工具软件方面都有较大的进步。在航天等领域，对软件的苛刻需求更是促进了形式化方法的进一步发展。

4.5 发展趋势与展望

从目前现状来看，容错计算机技术大多应用于需要高可靠性和高安全性的特殊工业控制领域。但随着技术的不断进步，人力成本越来越高、人们生活品质不断提升，各种无人值守的应用场景会越来越多，机器代替人进行劳动已成为一种趋势，这种高可靠性和高安全性的需求必将深入我们的生活。与此同时，危及安全的工业控制领域也将对计算机产生新的更高的要求。在这种情况下，控制计算机系统的可靠性和安全性将成为一个越来越被关注的问题，高可靠容错工业控制计算机将是一个重要的研究和发展方向。

在硬件可靠性方面，可重构技术不仅能够在计算机发生故障时通过重构避免系统失效，而且能够使系统具有多种硬件功能，以适用于不同的实际需求；甚至在重构过程中采用智能方法，使得工业控制计算机具有自学习、自修复和自适应的功能。可以预计，可重构与人工智能是高可靠和高安全工业控制计算机的发展趋势。

未来对软件的可靠性和安全性要求也越来越高，软件引起的错误往往是灾难性的，因此软件的可靠性保障也是未来发展的一个重要方向。目前，形式化验证是保障软件可

信、可靠的一种颇具前景的手段。但这种方法还处于理论研究阶段，并未大量在实际工程中应用，采用形式化验证方法保障软件的可靠性和安全性还有很多工作要做。

5 结束语

经过几十年的发展，工业控制计算机已经成为工业控制领域不可或缺的重要组成部分，工业控制计算机的技术发展呈现多领域、交叉学科的特性，其研究领域涉及控制领域、电子学领域、计算机技术领域、网络技术领域、有线和无线通信领域，甚至生物学领域。工业控制计算机技术向着高性能、低功耗、高实时性、高可靠性、分布式、并行化、网络化、无线化、智能化的方向发展。工业控制计算机与当前计算机前沿技术越来越融合在一起，相互促进，共同发展。工业控制的应用领域也为计算机技术研究提供强大的驱动力和广阔的应用背景。可以预见，计算机技术的高速发展必将为工业控制计算机提供更为广阔和坚实的理论基础，工业控制计算机必将向着性能更高、更为可靠和安全、环境适应性更强、更为智能化的方向发展。

参考文献

- [1] GB/T 26802. 1. 工业控制计算机系统 通用规范[S].
- [2] GB/T 26803. 1. 工业控制计算机系统 总线[S].
- [3] 刘鑫. 工业控制计算机技术的最新进展[J]. PLC&FA, 2008(1).
- [4] 刘鑫,杨孟飞.“十一五”期间工业控制计算机技术发展探讨. 自动化博览,2007(3).
- [5] PICMG. PICMG CPCI-S CompactPCI Serial Rev 1.0. 2011[S].
- [6] PICMG. PICMG 3.0 AdvancedTCA Base Specification Rev 3.0. 2008[S].
- [7] PICMG. PICMG MTCA. 0 MicroTCA Rev 1.0. 2006[S].
- [8] PICMG. PICMG 2.0 R3.0 CompactPCI Specification. 1999[S].
- [9] PC/104. PC/104 Specification Version 2.5[S].
- [10] PC/104. PC/104-Plus Specification Version 2.2[S].
- [11] PCISIG. PCI Express Base 3.0 Specification[S].
- [12] PXISA. PXI Specification Rev 2.0. 2000[S].
- [13] Tom Williams. VME Turns 25... and Time Marches On. RTC. 2006.
- [14] Joe Pavlat. Introducing the 3rd annual CompactPCI and AdvancedTCA Systems Resource Guide [J]. CompactPCI and AdvancedTCA Systems. 2006.
- [15] PICMG. PICMG Specification MTCA. 0 R1. 0- Micro Telecommunications Computing Architecture Base Specification[S]. 2006.
- [16] Stuart Jamieson. MicroTCA offers a direct solution for tight cost and size constraint applications [J]. CompactPCI and AdvancedTCA Systems. 2006.
- [17] Andrew Brown. CompactPCI Express: Protecting CompactPCI investments made over the last 10 years[J].

- CompactPCI and AdvancedTCA Systems. 2005.
- [18] PXISA. PXI Express Hardware Specification Rev 1.0. 2005 [S].
- [19] 刘鑫. 工业控制计算机技术进入第三代[J]. 世界仪表与自动化, 2005, 9(2).
- [20] 刘鑫. CompactPCI/PXI 关键技术的发展与应用[J]. 航天控制, 2004(3).
- [21] 刘鑫. FLASH MEMORY 及其在控制系统中的应用[J]. 电子技术应用, 1996(1).
- [22] 刘鑫. 机电一体化技术手册[M]. 2 版. 北京: 机械工业出版社, 1999.
- [23] 刘鑫. 32 位嵌入式 STD 总线工业控制机[J]. 电子技术应用, 1995(9).
- [24] ENSLO, PH. Ed. Multiprocessors and Parallel Processing[M]. New York: John Wiley. 1974.
- [25] Christopher P Bridges, Tanya Vladimirova. NASA/ESA Conference on Adaptive Hardware and Systems, 2008[C]. IEEE.
- [26] 杨毅, 张晓钟, 孙敏. ARM 处理器在我国 MID 领域的现状及发展趋势[J]. 微型机与应用, 2013, 31(23).
- [27] Wind River. Wind River VxWorks[OL]. <http://www.windriver.com/products/vxworks/>.
- [28] QNX Software Systems. QNX Neutrino RTOS[OL]. http://www.qnx.com/products/neutrino_rtos/.
- [29] RTEMS. Quick_Start[OL]. http://www.rtems.com/wiki/index.php/Quick_Start.
- [30] Marcus Völp. Prototypical Design and Implementation of L4-SMPMicrokernel Mechanisms[D]. Karlsruhe: TU Karlsruhe, 2002.
- [31] 董巍, 马云. 基于 ReWorks 操作系统的实时多任务程序设计[J]. 中国新技术新产品, 2013(9).
- [32] 张少林, 杨孟飞, 刘鸿瑾. 空间应用 SoC 研究现状简介[J]. 航天标准化, 2012(3).
- [33] 张蕴玉, 唐祖平, 胡修林. 基于 DeltaOS 的系统软件设计[J]. 微计算机信息, 2005, 21(23).
- [34] Algirdas Avizienis, George C Gilley, Francis P Mathur, et al. The STAR (self-testing and repairing) computer: an investigation of the theory and practice of fault-tolerant computer design [J]. IEEE transactions on computers, 1971, 20(11): 1312~1321.
- [35] 杨孟飞, 华更新, 冯眼君, 等. 航天器控制计算机容错技术[M]. 北京: 国防工业出版社, 2014.
- [36] John F Hanaway, Robert W Moorehead. Spaceshuttle avionics system[M]. Washington NASA, 1989.
- [37] S D' Angelo, C Metra, S Pastore, et al. Proceedings of 1998 IEEE International symposium on defect and fault tolerance in VLSI systems, 1998[C]. 233~240.
- [38] Callahan T J, Hauser J R, Wawraynek J. The Carp architecture and C compiler[J]. Computer, 2000, 3(4), 62-69.
- [39] Waingold E, Laylor M, Srikrishna, D, et al. Baring it all to Software: The Raw Machine[J]. Computer, 1997, 30(9), 86-93.
- [40] Goldstein S C, Sehnlt H, Moe M, et al, Proceedings of IEEE Symposium on Computer Architecture, 1999[C].
- [41] Kagotaniand H, Sehnlnt H, Proceedings of IEEE Symposium on FPGAs for Custom Computing Machines, 2003[C].
- [42] Sehnlt H, Whelihan D, Tsai A, et al. Proceedings. IEEE Custom Integrated Circuits Conference, 2002[C].
- [43] Yuan C, Pillai P. Proceedings. IEEE/ACM International Symposium on Microarchitecture, 2000[C].
- [44] Higuchi T, Iwata M, Takahashi E, et al. 26th Annual Conference of the IEEE Industrial Electronics Society, 2000[C].
- [45] Keymeulen D, Zebulum R S, Jin Y, et al. Fault-Tolerant Evolvable Hardware Using Field-Programmable Transistor Arrays[J]. IEEE Transactions on Reliability, 2000, 49(3): 305-316.

- [46] Forrest S, Perelson A S, Allen L, et al. Proceedings on 1994 IEEE Computer Society Symposium on Research in Security and Privacy, 1994[C].
- [47] Mario G. Perhimschi, Hever Moncayo, et al. Integrated Framework for Aircraft Sub- System Failure Detection, Identification, and Evaluation Based on the Artificial Immune System Paradigm[C]. AIAA Guidance, Navigation, and Control Conference, Chicago, Illinois, 2009[C].
- [48] Bowen, J, Stavridou, V. Safety-critical systems, formal methods and standards[J]. Software Engineering Journal, 1993, 8(4):189-209.
- [49] Jean Souyri, Virginie Wiels, David Delmas, et al. Formal Verification of Avionics Software Products[J]. FM 2009: Formal Methods, 2009(5850):532-546.
- [50] Esteve, M A, Katoen, J P, Nguyen, V Y, et al. 34th International Conference on Software Engineering 2012[C].

作者简介

刘 鑫 研究员，中国计算机学会工业控制计算机专委会主任，中国计算机行业协会副会长，中国自动化学会理事。



杨 桦 研究员，中国计算机学会工业控制计算机专委会副主任兼秘书长，长期从事星载容错计算机的体系结构、软硬件的设计与研究，负责完成了多个卫星平台星载计算机及软件型号研制任务、核高基国家重大专项等预先研究课题，目前已获得国家科技进步二等奖一项、北京市科技进步一等奖一项、部级以上科技进步一等奖、二等奖、三等奖各一项，型号首飞三等奖一项。



龚 健 高级工程师，中国计算机学会会员，从事星载容错计算机研制工作，研究方向是智能容错技术、高可靠嵌入式操作系统。



抗恶劣环境计算机技术的现状与发展趋势

CCF 抗恶劣环境计算机专委会

摘要

随着科学技术及信息化技术的迅猛发展，抗恶劣环境计算机应用已逐渐从军用领域拓展至冶金、石油勘探等民用领域，所需面临的环境也日益复杂，既需要应对极端恶劣的物理环境、太空环境及复杂的电磁环境，又需要面对日益开放的网络环境。环境的日益恶劣推动了抗恶劣环境计算机技术的不断发展。本报告从容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术及新材料与新工艺技术等方面出发，重点分析了国内外抗恶劣环境计算机技术的发展现状，总结了国内外研究进展差距，并结合我国的发展现状和不足，对未来抗恶劣环境计算机技术的发展给出了展望和预测。

关键词：恶劣环境，容错体系结构，加固防护技术，环境感知技术

Abstract

With the rapid development of science and information technology, the application of rugged computer has developed from military to civilian fields such as oil exploration and so on. The environment faced is more complex, for example, physical environment, space environment, network environment, and electromagnetic environment. The complex environment promotes the research of rugged computer. The paper puts important emphasis on analyze the development status of the technology of resistance to severe environment at home and abroad from the aspect of fault-tolerant architecture, software reliable technology, chip protected technology, Internet of things, environmental awareness technology, reinforced and protected technology, net electricity technology, and new material and new technology, then summarizes the gap between home and aboard, and comes to the conclusion of the prospect of the technology of resistance to severe environment.

Keywords: severe environment, fault-tolerant architecture, reinforced and protected technology, environmental awareness technology

1 引言

抗恶劣环境计算机主要是指能在恶劣环境下正常工作的计算机。传统的抗恶劣环境计算机主要应用于军用领域。随着军事作战环境的日益复杂，计算机的抗恶劣环境设计开始面向极端的物理环境、电磁环境和太空环境；随着网电空间的迅猛发展，开放的网络环境也对计算机的抗恶劣环境技术提出越来越高的要求；另外，随着无所不在的计算及物联网时代的到来，抗恶劣环境计算机在石油勘探、冶金、化工、地质勘探、矿产开

发、水文、气象、公安等领域得到了更为广泛的应用。因此，为满足未来纵深发展的需求，进一步研究抗恶劣环境计算机技术具有重要意义。

在军事、空天探测、野外作业及冶金、化工等领域的需求推动下，未来的抗恶劣环境计算机将向高性能、高可靠、轻型化、智能化、功能综合化方向发展，新的抗恶劣环境技术有待研究，如图1所示。在军事领域，随着信息战日益成为主要的战场模式，要求抗恶劣环境计算机不仅具有面向复杂恶劣环境的适应性，还应具备与其他处理设备协同工作的能力。如与传感器信息的融合，提高了网电空间战中作战态势的感知能力；在航空、航天领域，为进一步降低重离子对计算机的损害性，要求抗恶劣环境计算机具备抗重离子冲击产生的辐照效应能力，并遵循航天事业的节约型发展原则，采取低成本的集成电路设计；在冶金、石油勘探、矿产开发等工业领域，为提升操控作业的安全性及可靠性，要求抗恶劣环境计算机一方面能够智能化地进行排故、勘探、采集等操作，另一方面能够在抗恶劣环境的硬件平台上通过采取合理的软件可靠性措施提升整个计算机系统的稳定性。

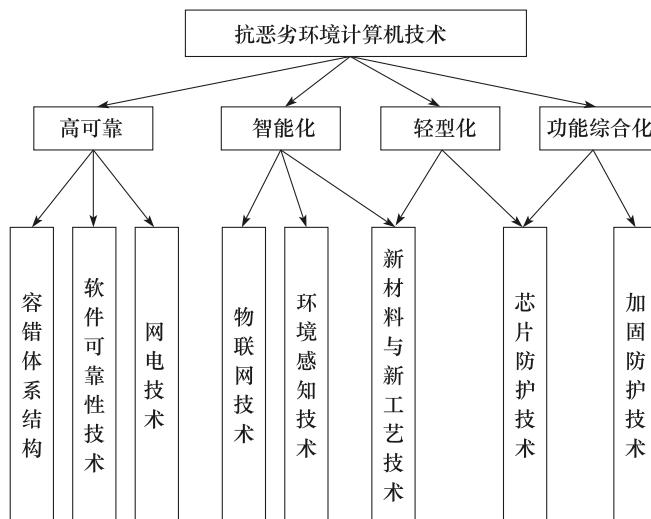


图1 抗恶劣环境计算机技术

综上所述，目前的抗恶劣环境计算机不但要具有面向复杂恶劣环境的适应能力，还需迎合军用、民用领域对抗恶劣环境计算机的深层发展需求，具有高可靠、小型化、轻型化、智能化等特性，所涉及的技术包括容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术、新材料与新工艺技术等。基于上述技术，进一步展开抗恶劣环境计算机技术的深入研究，不但能够提升计算机的多维环境适应能力，还可以拓宽抗恶劣环境计算机的应用领域，推动抗恶劣环境计算机的进一步发展。本报告从国内外抗恶劣环境计算机技术的各个方面进行技术发展现状和趋势的讨论，并就我国抗恶劣环境计算机技术发展提出有益的建议和发展望。

2 国际研究现状

国外的抗恶劣环境计算机技术主要涉及容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术及新材料与新工艺技术。

2.1 容错体系结构

在抗恶劣环境计算机领域，计算机体系结构的研究主要集中于容错体系结构^[1,2]，其研究热点正从单机向分布式系统方向发展，代表公司有美国容错电脑公司（STRATUS）及欧洲 GUARDS。

作为国际知名的容错计算机厂商，STRATUS 重点研究高性能分布式容错系统，即通过通用微处理器及微计算机，采用在局部网络中注入全局管理、并行操作、自治控制、冗余和错误处理等来实现其可靠性。其推出的 ftServer 系列容错服务器，体系结构采用 DMR/TMR（双模冗余/三模冗余）硬件容错架构，如图 2 所示。在 DMR 模式下，系统由两块主板、两套 IO 子系统组成。每个主板上可装配 1~2 个微处理器，支持对称多处理（SMP），容错总线隔离了微处理器与 I/O 的直接连接，两套主板及全部冗余部件构成具有一个软件映像的逻辑计算机，其中，两套主板锁步运行同样的应用、板级错误检测，专有的计算法则及其他诊断及逻辑功能放在监控系统中。当检测到处理器错误或部件故障时，故障部件被自动隔离，其他系统不受影响地连续运行，不需中断处理，性能及数据完整性没有任何损失，对用户及应用而言，故障是透明的。在 TMR 架构中，增加了第三块容错主板，提供了额外的可靠性，可以提供超过 99.999 + % 的可用性。假设故障发生，第三块主板被隔离，系统缺省进入 DMR 模式，仍具备全冗余的 99.999 + % 可用性。

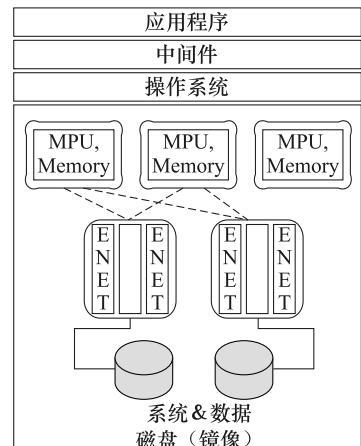


图 2 ftServer 体系结构

欧洲 GUARDS 研制的容错计算机则采用商用部件（COTS 技术）、开放结构、多层次的容错技术，通过组合，满足对计算机的高可靠性需求。该容错体系结构有较明显的性价比优势，大量采用 COTS 技术，可以在不同方向上对系统进行扩展，升级能力好。

可以看出，国外在容错体系结构的主要目标是利用组件化设计思想，通过开放式、高安全的体系结构设计，采用软硬件的可靠性技术提高组件的安全性和可靠性，以构建系统的高安全性，满足在恶劣环境下计算机系统安全、可靠运行的需求。

2.2 软件可靠性技术

在国防军事、航空航天、工业控制、金融等对可靠性和安全性有着极高要求的领域中，由于软件失效和安全性问题造成的灾难性事故层出不穷，对软件的可靠性分析、评估和保障变得尤为重要。早在 20 世纪 80 年代中期，西方各主要工业强国就在可靠性与安全性方面确立了专门的研究计划和课题，如英国的 AIVEY（软件可靠性和度量标准）计划、欧洲的 ESPRIT（欧洲信息技术研究与发展战略）计划、SPMMS（软件生产和维护管理与保障）课题、Eureka（尤里卡）计划等^[3]。每年，各国对此都有相当比例的投入。计算机可靠性的需求推动了软件可靠性技术的不断发展。目前，国外对软件可靠性技术的研究主要以软件失效模式分析与建模技术为主，并推出了相关系列产品。如美国的 ITEM 公司开发的 FMEA 软件、Relex 可靠性系列软件、瑞蓝公司开发的 XFMEA 软件等。虽然这些软件在功能上存在差别，但其主体思想都是通过识别软件存在潜在的失效模式、评估失效模式对系统的影响，找出软件失效的原因，并提出有效的措施或修改意见。此外，国外还加强对软件故障注入技术的研究，美国的 California、CMU、Duke、IBM、Illinois、Michigan、NASA、RST、Tandem、TAMU、Virginia 等大学和研究部门以及法国的 LAAS-CNRS 等大学和研究部门的故障注入研究工作都处于国际领先地位，其成果集中体现在对故障注入具体实现方法的改进和创新上，包括对在不同抽象层次上注入故障的方法、尽可能精确模拟真实故障影响、减少试验开销和提高结果精确性、构造实用的故障注入工具等方面的研究，其中，许多实用的工具已投入使用^[4-14]。

可以看出，国外对于抗恶劣环境计算机软件可靠性技术的研究主要集中在软件故障注入技术及软件失效模式分析与建模技术上，并努力提升产品的实用化水平。

2.3 芯片防护技术

恶劣的空间辐射环境，促进了国际学者对芯片防护技术的不断研究^[15-17]。目前，国外抗恶劣环境计算机芯片防护技术的发展主要集中于 SOI(Silicon on Insulator)。相较于体硅技术，SOI 电路的抗辐照能力提高了 100 倍。通过表 1 的展示，可以看出与标准体硅技术相比，SOI 技术电源电压更低，总剂量辐射不相上下（SOI 也需要抗总剂量辐射加固），剂量率翻转指标大很多，抗单粒子翻转能力更强，无闩锁效应。

表 1 SOI 技术与体硅技术的抗辐照特性比较

工艺	电源电压/V	总剂量/rad(Si)	剂量率翻转/(rad(Si)/s)	单粒子翻转阈值/MeV cm ² /mg	闩锁
CMOS6R (体硅 CMOS)	5.0	<1M	109	40	有
CMOS7 (SOI 技术)	3.3	>1M	1 011	>40	无

在抗辐照 SOI CMOS 集成电路制造方面，国外以美国的 Thomson CSF 公司和 Honeywell 公司、AMD 公司为主。Thomson CSF 公司针对军事与空间抗辐照应用，开发了商品化的 CMOS/SOI 电路，包括 SRAM、A/D 转换器等^[18]。Honeywell 公司商品化的 HX6156 系列产品主要用于航空航天及军工电子领域，其抗辐照总剂量水平达到 1 Mrad，在 3.3V 工作电压下其功耗为 0.14mW/Gate/MHz，当电压为 2.5V 时其功耗为 0.08mW/Gate/MHz。除了具备强的抗辐照性能，SOI 技术还具备低功耗特性。在 AMD 公布的一款 SOI 芯片的测试结果中，相较于传统的体硅工艺设计，该芯片显示出最高可达 40% 的功耗节省的可能性。此次发布的结果证实了在高性能设备上设计低功耗处理器时，SOI 是一项可取代传统体效应工艺的可行技术。此外，国外某些公司还开展 SOS（蓝宝石外延硅）技术及 GaAs 技术研究。SOS 材料同样具备理想的抗辐射能力和低功耗特性，但是其晶片易碎，成品率低，成本高，只能用于特殊的辐照环境。GaAs 电路的速度比其他材料高 5~10 倍，抗总剂量辐照能力强，被广泛应用于军事领域，是激光制导导弹的重要材料，曾在海湾战争中大显神威。但 GaAs 成本高，其单晶片的价格大约相当于同尺寸硅单晶片的 20 至 30 倍，另外 GaAs 材料导热性差，不适宜制作大功率器件。

在国际航空航天领域，国外对芯片防护技术的研究重点放在抗辐照能力强、功耗低、抗干扰能力强、集成密度高的新材料技术研究上，以满足设备抗恶劣太空环境日益增长的需求。

2.4 物联网技术

物联网具有泛在感知、可靠传送、智能处理等特点，其体系结构^[19]如图 3 所示，感知层、网络层和应用层的通信，可实现抗恶劣环境计算机的智能化方向发展。国外对物联网技术的发展多集中在安全技术领域。如 Mulligan^[20]等人总结和分析了物联网的国内外研究现状，并进一步讨论和展望了物联网安全和隐私问题；Medaglia^[21]等人总结和讨论了物联网面临的隐私与安全保护问题以及将来可能会面临的安全隐患；Leusse^[22]等人总结了一个物联网服务安全模型，同时分析了其各个模块的功能和作用；Hamad^[23]等人认真分析了电池能量消耗等资源消耗，总结了目前已有的物联网安全加密算法。RFID 作为物联网的基础技术，其安全技术研究也逐渐成为国外学者的研究热点^[24-32]。针对 RFID 的安全威胁主要有物理攻击、假冒攻击、重传攻击、追踪及窃听。为增强主动作为，国外主要集中在 RFID 的攻击技术上。2005 年 7 月，在美国拉斯维加斯召开的第 13 届国际安全会议上，研究者进行了一次演示实验，实验人员在距离射频识别阅读器 69 英尺远的地方，运用 RFID 攻击技术接收到射频识别阅读器的电磁波信号，而这个演示系统的最大设计阅读距离不超过 10 英尺。2006 年 2 月，以色列威兹曼大学的计算机科学教授 Adi Shamir 宣布，他能用一个极化天线和一个示波器来监控射频识别系统电磁波的能量水平。他指出，可以根据射频识别波瓣场强的变化来确定系统接收和发送加密数据的时间。根据这些信息，射频识别系统安全攻击者可以对射频识别的散列加密算法 1 (Secure Hashing Algorithm 1, SHA-1) 进行攻击，而这种散列算法在某些射频识别系统中是经常使用的。

按照 Shamir 教授的研究成果，普通的移动电话就会对特定应用场合的射频识别系统导致安全危害。2006 年 8 月，荷兰阿姆斯特丹自由大学的一个研究小组研究成功了一种称为概念验证（Proof of Concept, POC）的射频识别蠕虫病毒。这个研究小组在射频识别芯片的可写入内存中注入了这种病毒程序。当芯片被阅读器唤醒并进行通信时，病毒通过芯片最后到达后台数据库，而感染了病毒的后台数据库又可以感染更多的标签。该研究课题采用了包括 SQL 和缓冲区溢出攻击（Buffer Overflow Attack）等在内的常用服务器攻击方法。射频识别系统的标签数据可在一定距离内传输的特性，为攻击者的侦听（Sniffing）和数据欺骗（Spoofing）提供了方便。而且，即使射频识别系统的电磁波场强很小，电磁波传输的距离也是系统设计的最大阅读距离的很多倍。

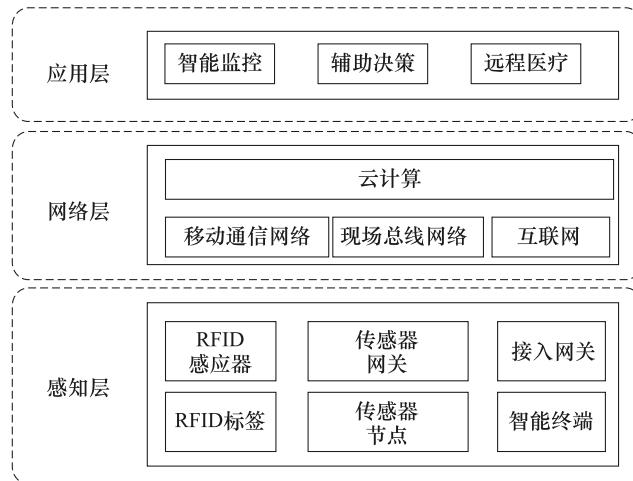


图 3 物联网体系结构

综上所述，美国“智慧地球”概念及欧盟“欧洲物联网行动计划”的提出，在世界范围对物联网的研究、应用和产业发展等起到积极的引导和推动作用。物联网的安全问题显得尤为重要，国外在大力发展物联网技术的同时，也越来越重视物联网安全技术的研究，并努力将其产品进行实用化推广。

2.5 环境感知技术

环境感知技术可实现计算机在恶劣环境下对目标的感知、识别与判断能力，有助于提高计算机恶劣环境的适应性。国外对环境感知技术研究的热点主要是多传感器技术。

近年来，谷歌、三星、苹果等巨头竞争研发的众多移动设备都采用了多传感器技术，以实现运动跟踪、数据收集、信息传输、环境感知等基本功能，使“人 - 设备 - 环境”间完成信息互动，其技术方案如图 4 所示。其中，三轴陀螺仪、加速传感器和距离传感器能够实时跟踪身体运动，环境声传感器（麦克风）能够有效输入并监测声音，GPS 能够实时监测用户地理信息，温度传感器能够实时监测环境温度，光传感器（摄像头）能够有效识别二维码、人脸信息等。

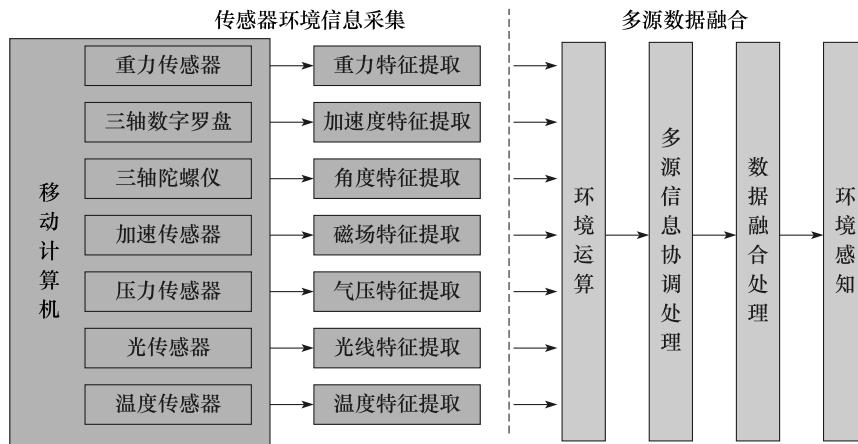


图 4 环境感知技术方案

在军事领域^[3-41]，环境感知技术多用于无人地面车辆（UGV）、自主地面移动平台（ALMP）、自主地面车辆（ALV）等无人机动平台应用，通过环境感知提供的世界环境模型、运动状态和定位信息保证其行驶的安全、稳定。在 UGeV 计划中，由卡内基 - 梅隆大学（CMU）国家机器人工程中心（NREC）研发的应用于大范围复杂越野路面的下一代无人高机动平台 Crusher，就是借助车载传感器技术检测从地图数据无法辨别的各种路面的行驶要求。值得注意的是，Crusher 无人机动平台在极端恶劣环境下具有较好的燃油经济性、较高的生存能力和较强的承载能力，展现了下一代军用无人机动平台对各种类型越野路面突出的适应能力和机动性能。除上述平台外，美国现役或在研的地面无人机动平台还包括洛克希德·马丁公司的 AMAS 平台、SMSS 平台和 MULE ARV A(L) 平台，上述平台均采用了环境感知技术^[42-46]。

随着设备向小型化、轻型化、智能化方向发展，微传感技术成为新的研究热点，在航空航天、生物医学方面得到了广泛的应用。例如，DARPA 研制的独立芯片级惯性导航和精确制导系统部件，比传统惯性器件尺寸更小、重量更轻、功耗更低，其工作功率不超过几十毫瓦，可降低武器系统对 GPS 系统的依赖。

此外，随着计算设备拥有更强的处理能力、更好的连接性以及创新的感应功能，国外知名公司已不满足于目前各种基于传感器的简单应用，而着力于新“环境感知”计算技术的开发。通过新环境感知技术，设备的软硬件传感器的多源融合可实现用户需求的预见，并为用户提供建议和引导，对促进计算机设备的智能化有着革命性的影响。目前，英特尔公司实验室已成功进行了很多演示，其 Socially Enabled Services (SENS) 项目能够感应并了解用户的实时活动；英特尔公司还与 Fodor's Travel 公司联合推出了 Personal Vacation Assistant 试验版，该成果可通过用户的当前位置以及形成信息等各种环境资源，为用户提供实时活动推荐。上述研究成果都对未来计算机的发展有着深刻的影响。

综上所述，在航空航天领域，国外的环境感知技术已实现了计算机的微型化及智能化，并在无人机动平台中有了成熟应用。新环境感知计算机技术的研究，将带动抗恶劣环境计算机向智能化方向发展。

2.6 加固防护技术

加固防护技术是实现抗恶劣环境计算机系统稳定性和安全性的重要途径，国外的加固防护设计研究主要集中于热设计技术、工程化仿真技术^[47-53]等。

在热设计方面，为满足未来机、舰、车载设备的散热设计要求，国外重点研究液冷技术及射流冷却技术^[54]。DHY 生产的液冷冷板和机架在 F/A-18 和 F-22 飞机上已经应用。在射流冷却技术方面，以亚利桑那大学和马里兰大学为首的研究机构在射流冷却换热理论上进行了深入的研究。ISR、Parker 等公司可提供部分采用射流冷却技术的产品。为 F22 提供集成机架和 LRM 模块的 DHY 公司拟在未来机载电子设备上采用射流冷却技术。

在抗恶劣环境计算机领域，工程化仿真技术^[54]主要应用于雷电及 HIRF 防护方面。商用运输飞机主要根据数值仿真的计算结果进行 FAA/JAA 防雷击适航认证。工程仿真技术用于验证全机雷击试验，为飞机局部电磁脉冲试验计算电缆电流和端部电压，在取得大量仿真和局部试验数据后，不再需要进行全机雷击试验。国外 EMA3D 公司利用软件进行工程仿真，相对于全机雷击试验，软件仿真能节省研制经费 150 万美元。在航空领域，工程仿真技术还可以应用于仿真和测试在 HIRF 电磁环境条件下，机舱内部传导干扰和辐射干扰的情况，并在波音 787 商用喷气式飞机闪电/HIRF 得到了论证。

可以看出，国外在抗恶劣环境计算机加固防护技术的研究上，目标非常明确，就是通过优化的物理工程手段，帮助计算机对抗大气及外太空中的恶劣环境条件。

2.7 网电技术

信息技术的发展，催生了第五作战领域——网电空间的形成。欧美国家围绕网电空间的军事活动日益频繁，如组建网电空间司令部、开展“网电风暴”演习等。其对网电技术的应用主要体现在三个方面：对敌方指挥体系的破坏，对敌方信息的误导以及上述两种方式的混合破坏^[55-59]。

对敌方指挥体系的破坏主要通过发送电脑病毒、逻辑炸弹等方法破坏敌方网电系统，造成敌方指挥控制系统的瘫痪，或侵入敌方网络系统，窃取敌方的保密文件，删除、修改敌方系统中的数据，释放计算机病毒，埋藏后门程序等。美军为提升自己的网络攻击能力，已成功研制了包含 Stuxnet 在内的两千多种病毒，并利用 Stuxnet 病毒对伊朗铀浓缩设施实施了长达 3 年的攻击。另外，美军方正研制、配发用以专门破坏敌军 C3I 系统的网络空间武器，这些网络空间武器专门破坏敌方指令，甚至能潜入敌方的侦察卫星、导航、防空和信息通信指挥系统。其中，舒特系统已经在入侵、窃取、注入、控制等方面取得了很大成功。目前，舒特系统已经实验了 6 代，形成了舒特 I~VI，并不断改进其侦察能力、欺骗能力、软件算法能力、评估能力等，其作战对象也在不断扩展：舒特 I 主要针对俄制防空导弹系统，舒特 II、III、IV 进一步拓展到机动战术弹道导弹系统、一

体化防空系统和指挥控制网络、战略导弹指挥控制网络，舒特 V、VI 已扩展到国家级指挥控制系统、一体化战略防空系统及反导反卫系统，可进一步扩展至通信系统等战场信息系统。

对敌方信息的误导主要通过向敌方网电系统传输假情报，改变敌方网电系统功能、对敌决策与指挥控制来产生信息误导和流程误导。当敌方使用无线信道开始传输信息时，网电侦察设备就可以截获信息，并做相关的解密处理，在掌握其网电协议及密码后，就可通过无线的方式将更改后的信息注入敌方网电之中。

混合破坏主要通过综合利用体系破坏和信息误导，并与其他信息作战样式相结合，对敌方指挥控制系统造成多重杀伤功效。例如，利用强烈的干扰信号去压制战场网电无线信道，可以达到扰乱战场网电正常工作的目的。随着美国军事行动对 GPS 系统的依赖性逐步增强，为抵抗他国的混合破坏行为，美军开始重视发展 GPS 系统的安全与防护措施，大力开发抗干扰技术、研制“卫星受威胁与攻击告警”（STW/AR）系统等。

综上所述，空间已成为高技术战争的制高点。国外的网电空间技术已取得蓬勃的发展，较之欧美，我国的网电技术研究仍有较大的差距，在未来的网络一体化作战中势必缺乏竞争优势。

2.8 新材料与新工艺技术

在未来计算机系统小型化、集成化要求越来越高的大背景下，研究新材料、新工艺技术可实现计算机芯片的微型化、低功耗及集成化，并对未来武器系统小型化、智能化和轻量化产生颠覆性的影响。

近年来，国外发达国家加速对新材料及新工艺技术的研究。在新材料方面，美国已经建立了世界最大的 Material ConnecXion (MC 新材料图书馆)，该图书馆是世界唯一的国际化创新材料咨询服务机构，并且也是全球最大的创新材料集成与应用的图书馆。在欧洲，Material Xperience 由 Materia 公司于 2008 年在荷兰创建，并快速成为欧洲极具影响力的新型材料信息中心。同时，欧盟也建立了欧洲技术平台，其中涵盖了新材料发展各领域，主要目的是制定新材料相关路线图，并加强研究人员之间的合作和信息交流，最终增强欧洲科技竞争力。欧美国家的重视推动了新材料技术的飞速发展。如先进的 SiCp/Al 基等金属基复合材料已成功应用在航空航天等领域。美国海军飞行动力试验室研制了 SiCp/Al 基复合材料薄板并应用于新型舰载战斗机。俄罗斯航空航天部门将 SiCp/Al 基复合材料应用于卫星的惯导平台。

随着计算机对微型化及抗恶劣环境能力需求的日益增长，众多公司展开可取代硅晶片的新型材料研究，以满足设备电路的微型化、耐高温及高频率要求。如美国 IBM 公司采用石墨烯材料，成功研制出了首款由石墨烯圆片制成的集成电路。该集成电路由一些石墨烯场效应晶体管组成。其长度仅为 550 纳米，混频可达 10GHz，而且可以承受 125℃ 的高温。同时，美国电气与电子工程师学会（IEEE）出版的《IEEE 波普》杂志指出，该集成电路为一个宽频无线电频率混频器，其成果预示着由该集成电路制成的芯片，有

助于设备在无法接收信号的地方正常工作。

在新工艺方面，微系统集成方法与工艺有了新的突破，微电子器件特征尺寸继续减小。2014年，美国和日本先后研制了采用14纳米工艺实现的微处理器和现场可编程门阵列产品，以及15纳米工艺实现的闪存；2014年，三星公司14纳米三栅极FinFET芯片工厂开始批量生产。相比20纳米半导体工艺，14纳米工艺可以将处理器芯片的性能提升20%，功耗降低35%，占用面积减少15%。借助14纳米工艺，微处理的性能进一步提高的同时，功耗和成本也有所降低。DARPA在“高效线性全硅发射机集成电路”项目下成功研制出首个可工作在94GHz的全硅单片集成信号发射机系统级芯片，将原本由多个电路板、单独的金属屏蔽装置和多条输入/输出连线组成的发射机集成到了大约 1.3cm^2 左右的硅芯片上，实现了硅基射频器件输出频率和效率的大幅提升。此项技术有望为未来军用射频系统提供新的设计架构，使下一代军用射频通信系统体积更小、重量更轻、成本更低、功能更强。2013年，DARPA还开发出二维光学相控阵芯片，将4096个纳米天线集成到一个硅基底上，尺寸可媲美针尖，该芯片的成功表明异质、异构硅基光学集成技术取得重要进展。

可以看出，国外高度重视新材料、新工艺技术的研究，其产品正从芯片级、组部件级向复杂程度更高的系统级等发展，未来将广泛应用于军事领域中。

3 国内研究进展

结合我国航空、航天、冶金、海洋、石油勘探等领域对抗恶劣环境计算机性能及环境适应能力提升的需求，国内抗恶劣环境计算机技术研究取得了长足的发展。

3.1 容错体系结构

在抗恶劣环境计算机技术领域，我国的抗恶劣环境计算机容错体系结构研究主要集中于航空航天领域。

北京控制工程研究所和中国科学院计算技术研究所^[60]研制了32位星载容错控制计算机系统及地面仿真空间环境综合验证平台，构建了动态可调整多级容错体系，创建了多级异构和冷热可变容错系统模型，并突破了多级动态可调整容错系统的同步-切换-重构技术，有效解决了空间飞行器资源有限而又面临多级系统长寿命与强实时之间相互制约的难题。本项目的成功是我国星载计算机领域的历史性突破，为我国航天领域重大专项成功实施发挥了不可替代的作用，具有重大战略意义。目前，项目成果已成功应用于80颗以上卫星，占同期国产卫星飞船的85%以上，包括嫦娥、载人飞船、通信、导航、科学试验、遥感等多种空间飞行器，其中最长已在轨工作七年半，累计在轨超过100余星年。中国航天科技集团公司组织的成果鉴定意见指出，星载计算机体系结构、多机动态可调整容错系统的同步-切换-重构、面向多故障和符合故障的诊断技术居于同期

国际领先水平。

随着二代导航、载人航天、深空探测等空间应用对低功耗和抗辐射容错能力提出更高的需求，可重构的容错体系结构研究及免疫与自愈相结合的综合容错处理技术成为国内众多研究机构的研究热点，但多处于理论研究和验证阶段。如航天 502 所的硬件进化重构容错研究，提出一种适用于空间应用领域进化容错的 FPGA 故障模型和故障检测方法，并验证了进化重构容错的可行性。尚利宏等提出了一种基于二阶近似域划分的可重构容错片上系统，验证了在商用 FPGA 上构建高可靠、低时延的复杂系统的可行性^[48]。文献 [61] 研究了一种基于动态可重构的容错体系结构，即通过基于系统降级的重构策略来实现系统级容错，并采用 LEON2 作为处理单元，对容错模块功能进行了仿真验证，结果表明容错控制满足预期的设计需求。针对于免疫与自愈相结合的综合容错处理技术研究，装备指挥技术学院在分布式卫星系统的容错设计中引入了“免疫”和“自愈”概念，“免疫”是指系统具有防范某些故障发生的能力，“自愈”是指系统具有监测、发现、容忍、处理故障和恢复正常的能力。其具体处理办法是星载计算机硬件平台采用基于温备的系统级双机容错方案：单节点具有两台工作机组，当两台机组均正常工作时，其中一台机组作为主机，另外一台机组作为备机，一旦主机发生故障无法完成指定的任务，原来的备机将成为新的主机，并接管当前的工作。对硬件平台的系统监控采用硬件与软件相结合的容错方法，并采用内部和外部 watchdog 监控策略。内部 watchdog 监控利用 AT91RM9200 处理器内部软件 watchdog 功能实现，外部 watchdog 监控利用接口板中的 FPGA 和专用 watchdog 芯片实现。由 AT91RM9200 内部软件 watchdog 和接口板上的硬件 watchdog 构成了硬件平台中的两级监控、一级冗余的 watchdog 监控机制^[62]。

综上，我国的容错体系结构多用于航空航天领域，如何开展深入研究以拓展至其他应用领域，将是抗恶劣环境计算机容错体系结构技术未来的发展方向。

3.2 软件可靠性技术

目前，我国成立了很多软件测试中心，如中国航天可靠性与安全性研究中心、武汉大学软件工程国家重点实验室、北京航空航天大学可靠性工程研究所、同济大学安全软件测试评估研究所等。但我国的可靠性关键领域的相关研究起步较晚，软件可靠性设计水平还比较低，尽管采取了一些措施，但大多数还是基于经验的、个别的、鼓励的措施，难以对系统进行深入的设计和管理，且基本集中在军事领域。

在军用领域，为提高我军武器装备复杂电磁环境下的作战能力，针对武器装备软件在复杂电磁环境（由空域、时域、频域、能量上分布的数量繁多、样式复杂、密集重叠、动态交迭的电磁信号构成）下量化测试验证和评估需求，围绕复杂电磁环境下武器装备高效、低成本软件测试与评估目标，以《武器装备电磁环境适应性试验技术指南（1.0 版）》为基础，我国面向防空反 X 武器装备弹上的导引头系统、雷达系统等，重点研究复杂电磁环境下的软件失效模式分析及建模技术、复杂电磁环境下的软件故障注入技术、复杂电磁环境下的软件适应性评价技术，形成武器装备复杂电磁环境下软件测试评估平

台，为武器装备在复杂电磁环境下软件高效、量化测试提供技术手段。其中，针对软件失效模式分析的安全标准和指南等相关内容，已纳入 GJB-1391—2006^[63]。

此外，我国许多高校及研究所也着重开展民用软件故障注入技术的研究，如清华大学、北京航空航天大学、航天部 502 所、航天部 771 所，以及中科院上海微系统与信息技术研究所等，其研究方法主要是通过修改系统内存或者寄存器的内容来注入故障，或者使用程序变异来实现故障注入^[64-69]。

3.3 芯片防护技术

随着我国空间技术的迅猛发展，我国加强了芯片防护技术研究。国内一些研究机构展开了 SOI 材料和 SOI 电路的研究工作，并取得了不错的成果。

在 SOI 研究方面，国内首次研制成功的航天用 SOI 工艺 16 位微处理器 1750A^[70]，LET 阈值 $L_0 = 67 \text{ MeV cm}^2/\text{mg}$ ，属于抗单粒子效应能力较强的器件；国产 128K SOI SRAM^[71]，其翻转 LET 阈值大于 $61.8 \text{ MeV cm}^2/\text{mg}$ ；中国科学院微电子研究所承担的国家攻关项目“亚微米 CMOS/SIMOX 器件和电路的研究”，开展了部分耗尽（PD）和薄膜全耗尽（FD）SOI CMOS 器件工艺和相关电路的研制，研制成功 $0.8 \mu\text{m}$ 全耗尽 SOI 101 级环形振荡器，在 3V 电压下门延迟为 69ps/门；研制出 SOI/CMOS 64kbit 静态随机存储器，抗 Y 总剂量达到 1Mrad(Si)，抗单粒 LET 值大于 $59 \text{ MeV cm}^2/\text{mg}$ （无翻转）。另外，上海新傲科技有限公司可提供商业化的 SOI 材料。

除此之外，由于 SoC 新型芯片集成技术和 MEMS 能够满足航天电子要求的集成化、小型化、高性能、低功耗要求，我国也展开了 MEMS（微电机系统）和 ASIC（专用集成电路）相结合的 SoC 新型芯片集成技术研究。目前，国内集成电路及 MEMS 相关基础产业环境薄弱，航天微电子技术和产业远远滞后于系统的应用需求，航天电子系统大量使用商业器件且高端芯片大都依靠进口，商用集成电路和元器件无法完全满足军事及航天应用的需求，且存在重大技术、安全隐患。因此，虽然我国在航天专用集成电路技术和产品的研制方面取得了一定的成果，已经具备一定的研制能力，但尚有很多关键技术有待突破。

可以看出，SOI 结构能有效地克服体硅材料的不足，充分发挥硅集成技术的潜力，在高性能、高压、高温、抗辐照、低功耗等领域均有极其广泛的应用。加强我国 SOI 技术研究，将进一步提升我国计算机设备的恶劣辐照环境的适应能力。同时，进行航天专用集成电路及与 MEMS 相结合的芯片集成技术研究，研制航天专用的自主芯片以降低成本，也将推动节约型航天事业及航天自主可控发展。

3.4 物联网技术

随着物联网技术的广泛应用，我国的物联网技术研究有了长足的发展，并在军事领域取得了丰富的成果。在抗恶劣环境计算机领域，国内的物联网技术研究主要集中于

RFID 技术研究及节能技术研究。其中, RFID 技术研究主要针对军事领域, 并有了成熟应用; 节能技术研究主要针对民用领域, 目前, 多数仅限于实验验证阶段, 尚未推出成熟的产品应用。

在军用领域, 我国积累了丰富的射频识别 (RFID) 技术和产品研发经验。中国航天科工集团某研究所承担了我国军用射频识别标准验证项目工作, 对军用射频识别与空中接口协议标准进行了模拟仿真验证与半实物验证, 建成了射频识别软件仿真与硬件测试平台, 并设计了军用射频识别信息服务体系仿真验证平台。结合物联网安全技术的发展, 同时开展了针对芯片、协议、读写器等 RFID 系统各个重要组件的安全建模和攻击技术研究。

在民用领域, 我国高度重视物联网节能环保产业的发展, 并出台《关于加快发展节能环保产业的意见》(以下简称《意见》), 《意见》明确表示加强物联网节能技术创新, 提高技术水平。这一举措大力推进了我国各研究机构节能技术的研究。徐尉^[72]等研究了一种基于物联网/传感网的智能节能系统, 通过物联网的硬件及软硬平台设计, 实现系统的智能节能, 并通过实际应用环境平台的验证, 证明了智能节能系统的有效性和实用性。图 5 展示了采用物联网技术的智能节能系统结构。山东省计算机中心^[73]进行了基于物联网技术的数据中心动态节能研究, 研究中主要采用无线或 IP 网络的方式对数据中心的能耗情况进行采集, 并依据各环境的节能策略对能耗情况进行评估。本研究成果已在北京某银行总部的数据中心进行了测试验证。通过部署该节能系统, 能够有效分析数据中心各设备运行情况, 通过三维技术绘制温湿度云图, 发现温度或湿度热点。同时, 通过 IT 设备、网络设备、供配电系统数据的接入以及各设备耗电的采集, 可掌握能耗热点。

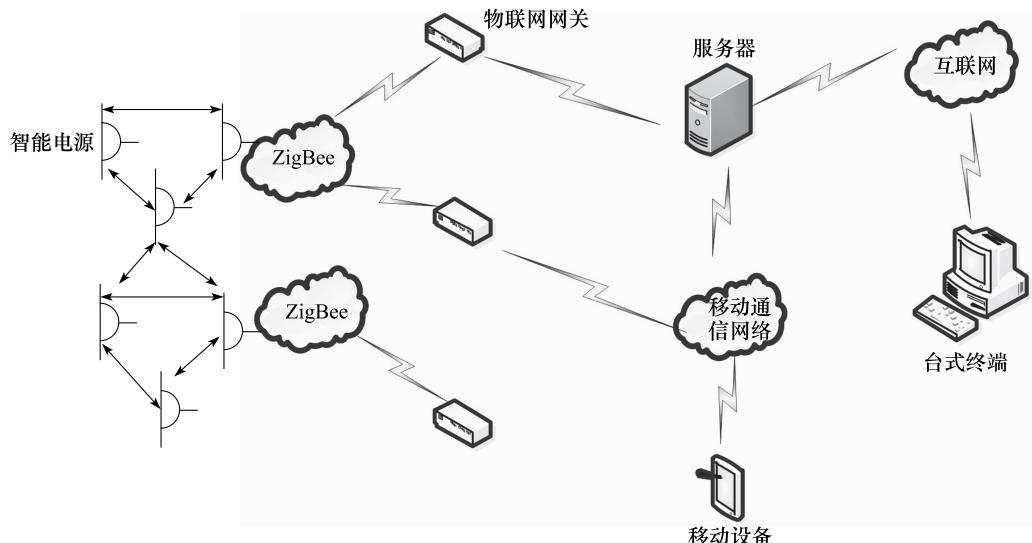


图 5 物联网环境下的智能节能系统结构

综上所述, 我国的物联网技术突飞猛进, 并在军用、民用领域有了丰硕的研究成果,

而如何将恶劣环境中物联网安全技术的研究成果向商用化转变及如何提升民用领域节能技术的实用化水平，对我国的研究学者将是新的机遇和挑战。

3.5 环境感知技术

在民用领域，我国的环境感知技术得到了广泛应用。百度、果壳电子公司等相继推出了成熟产品。果壳推出的移动设备拥有 WiFi 模块、蓝牙技术，并集成了陀螺仪、导航、加速度等多种传感器，可感知运动的状态等。

在军用领域，环境感知技术多用于无人机动平台的研究与开发^[74]。如中国兵器工业集团公司研发的混合动力无人地面车辆，是借助惯性导航元件、GPS 等传感器构成的可提供平台自身姿态、与环境相对位置或绝对位置信息的状态估计系统及通过各种雷达和相机等构成的能够提供世界环境模型的视觉系统，实现无人车辆的可靠运行，其基本原理如图 6 所示。其中，视觉系统主要用于障碍物检测、可通行区域提取、运动估计、地图创建等，其使用的传感器包括激光雷达、毫米波雷达、超声速雷达、单目相机、立体相机、全景相机、红外相机等。为了更好地控制地面无人机动平台的运动，准确的平台运动估计是必要的，特别是地面无人机动平台装备有悬挂系统，行驶环境较复杂，高速行驶在不平路面上时，必然会出现俯仰、侧倾、横滚、纵滑和侧滑等。无人机动平台的状态估计主要是融合多种传感器（如陀螺仪、GPS、加速度计、罗盘等）信息，估计平台的姿态、速度以及相对或绝对位置。通过卡尔曼滤波算法融合全球定位系统及导航系统，可在任何极端环境中提供长时间精确的定位信息。

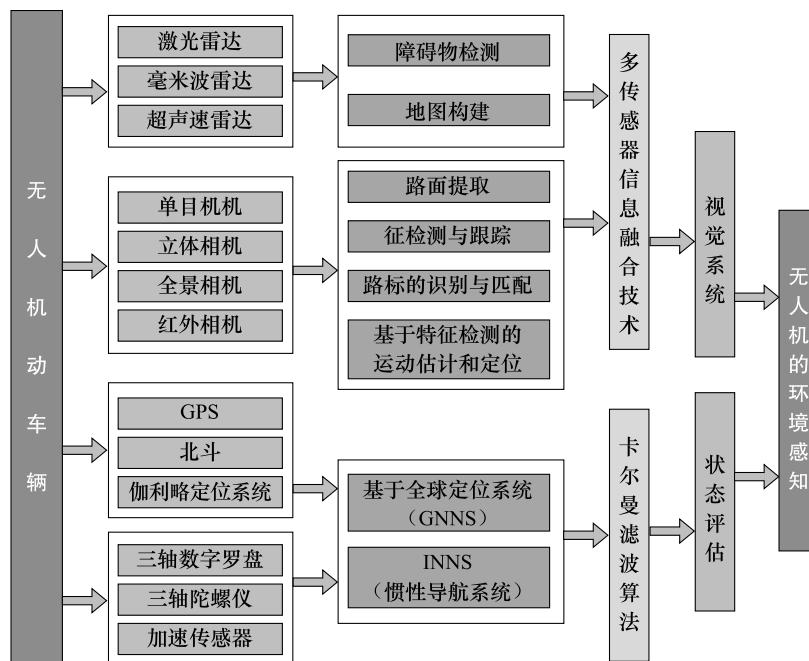


图 6 无人机动车辆环境感知技术方案

3.6 加固防护技术

我国的加固防护技术相对成熟，目前研究热点主要集中于热设计研究和电磁兼容性研究。

在热设计技术方面，我国颁布了关于电子设备热设计的统一标准，为设备的散热设计提供了基本的依据。近年，我国的热设计技术主要以热管技术、微通道技术、高导热绝缘材料、射流喷雾冷却技术、液晶屏低温快速加热技术研究为主^[50,51]。其中，热管技术已相对成熟，形成了风冷平板型散热器以及分离式热管、毛细泵回路热管、微型热管、径向热管等系列热管散热器，并广泛应用于航空航天、能源化工、电子设备冷却等领域。随着抗恶劣计算机向小型化、高性能、高可靠的方向发展，利用微尺度换热的特殊性达到高效冷却目的的微通道技术成为新的研究热点，此外，我国还开展高导热绝缘材料技术、射流喷雾冷却技术及加固液晶屏快速加热技术的研究，但都处于理论研究阶段。

在电磁兼容技术方面，我国着重开展板级及芯片级的电磁兼容仿真，但针对整机的电磁兼容仿真，尚处于理论研究阶段。

3.7 网电技术

从“十五”开始，我国相继设立网络安全对抗技术领域的系列预先研究课题，并成立了专业的专业技术组，指导、规划、统筹网电对抗领域的技术研究工作。在国家及军队政策的支持下，我国从IP网、固定电话网、网络电磁对抗及工业控制系统等方面开展了大量的信息安全对抗技术研究^[75-79]。

在IP网方面，通过多年的技术研究，以及“十五”和“十一五”预研课题的研制，我国突破了WiFi无线攻击技术。在固定电话网方面，突破了用户侦察技术、流量攻击技术等关键技术，研制了攻击样机，通过搭建的固定电话网仿真平台进行了试验验证。在网络电磁对抗方面，在电子对抗、雷达对抗、大规模网络攻击、对抗模拟演练等方面有了丰富的成果。在工业控制系统安全对抗技术和可信技术平台对抗技术方面，国内的研究主要集中在理论及方法层面，具体技术、工具研究较少。

相较于国外的网电对抗技术，我国与以美国为首的西方国家在系统与网络攻防装备和技术研究方面有着较大的差距。在国家经济社会发展、国防和军队建设以及各种军事行动对网电空间的依赖日益严重的今天，我国还需加强空地无线网络安全接入技术、空间网电态势感知技术、空间平台载荷可信安全保障技术等相关技术的研究。

3.8 新材料与新工艺技术

我国高度重视新材料、新工艺技术研究，并将其纳入战略新兴产业的重点扶持项目，已设立了与集成电路技术相关的国家重大科学专项、“973”和“863”等技术研发项目。

在新工艺方面，微纳米工艺有了重大的突破。中芯国际在国家“十一五”、“十二五”

重大专项的支持下，已实现 28 纳米工艺设计，并展开了 22/20 纳米技术的研究。22/20 纳米技术主要是以自主开发的 28 纳米工艺为基础，采用已掌握的“28 纳米成套工艺研究”项目的部分成果，根据 22/20 纳米世界产业主流设计规则和器件性能指标，引入若干个与 28 纳米工艺不同的工艺新模块，以此为基础确定 22/20 纳米的初始化工艺流程。

在新材料研究方面，军用新材料取得了巨大进展，如我国的某些航空航天领域新材料在单晶硅等半导体材料、超导材料及纳米材料等诸多领域均取得重大突破和快速发展，部分领域开始处于国际先进领先水平。在复合材料的研究上，主要集中于高体分比 SiCp/Al 复合材料、金刚石/Cu (Al) 复合材料等新材料技术研究，但尚处于试验验证阶段，暂无成熟应用。

从国内外计算机的抗恶劣技术的研究现状来看，抗恶劣环境计算机在容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术及新材料与新工艺技术等相关技术的推动下，其性能将会得到极大的提升，未来在军事、航空、航天、石油勘探等领域将会得到更为广泛的应用。

4 国内外研究进展比较

通过对抗恶劣环境计算机技术的国外研究现状及国内研究进展的分析，可以发现在抗恶劣环境计算的容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术、新材料与新工艺技术等方面，国内外研究进展各有不同，研究方向也略有区别。整体上，国外的抗恶劣环境计算机技术在软件可靠性技术、抗辐照技术、环境感知技术及网电技术方面的研究更先进、丰富，国内的相关研究起步较晚，距离国际领先水平有一定的差距。国内外研究进展对比如表 2 所示。

表 2 国内外研究进展对比

对比内容	国内研究进展	国外研究进展	备注
容错体系结构	拥有成熟的动态可调整多级容错体系、冗余体系结构及并行体系结构	采用组合化、开放式设计，并采用软硬件的可靠性技术提升系统可靠性	国内主要应用于航天航空领域
软件可靠性技术	针对专门领域开展基础理论与方法研究	软件失效模式分析技术及软件故障注入技术成熟	国外应用覆盖各个领域，国内多集中于军事领域
芯片防护技术	有自主 SOI 技术，SoC 新型芯片集成技术仍处于研究阶段	成熟的 SOI、SOS、GaAs 技术	与国外技术相差 1 代，大约 5~10 年
物联网技术	安全技术比较成熟，节能技术尚处于理论研究水平	物联网技术已应用于各个领域，尤其重视安全技术的研究，且成果丰硕	技术路线基本一致，国内仍需加强节能技术的实用化水平
环境感知技术	广泛用于民用领域，军用领域较少。微传感技术处于理论研究阶段	多传感器融合技术先进，军用、民用领域应用广泛。微传感技术研究比较深入，将逐渐应用于军事领域	技术手段基本一致，但在微传感技术研究方面与国外有一定差距

(续)

对比内容	国内研究进展	国外研究进展	备注
加固防护技术	热管技术成熟	射流冷却、液冷技术	国内液冷技术及射流冷却技术不成熟
网电技术	成立领导小组, 无验证和实验团队	建立有专门的验证和实验机构	国内跟踪国际
新材料与新工艺技术	复合材料相对落后, 初步实现 20 纳米新工艺	SiCp/Al 基复合材料、微纳米工艺成熟	新工艺技术相差 2 代, 大约 3~4 年; 复合材料处于验证阶段

5 发展趋势与展望

抗恶劣环境计算机技术不仅要具有面向更加恶劣、复杂环境的适应能力, 还需迎合军用、民用领域对抗恶劣环境计算机的深层发展需求, 这就驱动了抗恶劣环境计算机向小型化、轻型化、智能化、高可靠方向发展, 其抗恶劣环境技术的发展趋势如下:

(1) 容错体系结构

容错体系结构多用于航空航天领域, 满足未来航天任务对星载计算机的高性能、高集成化的要求, 这就需要研究可重构容错体系结构。通过硬件功能的在线重构算法, 可实现多个应用处理器和片上通信网络的配置及重构控制, 使系统能适应星载计算机在不同运行模式和应用场合下的应用需求, 实现计算机向小型化、高性能、智能化的方向发展。

(2) 软件可靠性技术

软件可靠性技术是保障抗恶劣环境计算机软件系统高效健康的关键技术。随着未来各军用、民用领域对软件系统高效健康开发的依赖性日益增加, 软件系统复杂化发展趋势下, 在严苛的应用背景中为保证软件系统健康, 应开展复杂软件系统健康因子体系建模与分析、健康状态实时监控与预警以及健康分析与度量技术的研究, 建立故障检测、恢复与安全隔离机制, 达到健康数据有效利用、软件系统健康全生命周期受控的目的, 以提升复杂软件系统设计、开发、测试、评估的健康保证能力, 为恶劣环境下各应用系统的可靠性运行提供高效支撑。

(3) 芯片防护技术

芯片防护技术主要涉及抗辐照技术及微系统化技术, 以满足抗恶劣环境计算机的高可靠性、小型化发展需求。目前, 我国已初步实现了依托 SOI 材料设计的微处理器等具有抗辐照能力的器件, 但为摆脱商业器件的高额成本及高端芯片受制于人的局面, 对芯片防护技术的研究需重点突破 MEMS (微机电系统) 和 ASIC (专用集成电路) 相结合的 SoC 新型芯片集成技术研究。

(4) 物联网技术

随着抗恶劣环境计算机向小型化、智能化方向发展，物联网技术可实现抗恶劣环境计算机的智能监控及辅助决策功能，尤其是在冶金、石油勘探、矿产开发等工业领域，物联网技术的智能化和节能化将有助于现场作业效率的提升，对节约成本也起到了积极的作用。目前，我国高度重视物联网节能技术和安全技术的研究，但还需进一步加强相关技术的探索，并实现安全技术的商用化。

(5) 环境感知技术

环境感知技术可实现计算机在恶劣环境下对目标的感知、识别与判断能力，有助于提高计算机恶劣环境的适应性。未来，随着抗恶劣环境计算机对智能化、微型化特性的需要及武器系统对作战环境态势感知能力的更高需求，环境感知技术将向微系统技术靠近，可将多种先进技术高度融合，将传统各自独立的信息获取、处理、命令执行等系统融为一体，以促进抗恶劣环境计算机的微型化和智能化，在军事应用中，更能促进武器信息系统性能的全面提高，对降低装备尺寸、重量和成本都有着革命性的影响。其涉及的核心技术有微传感器技术及微机电系统技术。

(6) 加固防护技术

加固防护技术是抗恶劣环境计算机技术的基础技术，已有几十年的研究历史。抗恶劣环境计算机的小型化与智能化，新材料与新工艺技术的应用，势必为加固防护技术研究带来新的挑战。在我国大力发展微纳米工艺及复合材料技术的同时，可随之展开高效散热技术及隔离密封技术研究。

(7) 网电技术

抗恶劣环境计算机不仅作为战机、战车、导弹以及军事综合电子信息系统的承载装备，还广泛应用于石油、冶金等国民经济关键行业领域，一旦遭受网络攻击，在军事领域不但会引起指挥控制的混乱，危及国家安全，在民用领域也会导致各信息系统的崩溃，给国家和个人带来巨大的经济损失。在信息安全的需求下，我国网电技术需立足自主创新，以抵近式及无线网络对抗技术、空地无线网络安全接入技术、空间网电态势感知技术、空间平台载荷可信安全保障技术、电磁安全对抗技术、工业控制系统安全对抗技术等为突破口，展开技术攻关，并构建典型的网电空间信息安全对抗模拟仿真验证环境，形成相应的网电空间信息安全对抗装备，实现工程应用，支撑新型武器装备与新兴产业的发展。

(8) 新材料与新工艺技术

新材料与新工艺技术可实现计算机芯片的微型化、低功耗及集成化，可对未来武器系统小型化、智能化和轻量化产生颠覆性的影响。未来，计算机将向微系统方向发展，研究微系统集成方法与工艺，将带动具备传感、处理、控制等多种功能的微系统快速发展，在大幅提升系统性能的同时，还能降低功耗及成本。未来的新材料技术，应大力发展战略金属基复合材料，以提高材料的可靠性、智能性，满足未来抗恶劣环境计算机的智能化发展需求。

6 结束语

随着科学技术的迅猛发展，抗恶劣环境计算机不但要面向极端的物理环境、太空环境、电磁环境，伴随网电空间的盛行，还需应对日益恶劣的网络环境。此外，随着抗恶劣环境计算机应用领域的不断拓展，对抗恶劣环境计算机的高可靠、智能化、小型化等提出了需求，在此背景下，国内外以容错体系结构、软件可靠性技术、芯片防护技术、物联网技术、环境感知技术、加固防护技术、网电技术、新材料与新工艺技术等为突破口，开展了新抗恶劣环境计算机技术的研究，并取得了丰硕的研究成果，不仅有效提升了计算机的性能及环境适应性，更促进了其在军事、航空、航天、冶金、化工、地质勘探、矿产开发等领域的更广泛应用。

依据国内外研究进展比较，国外在容错体系结构、软件失效模式分析技术、软件故障注入技术、抗辐照技术及网电技术等方面已相对成熟，而我国在相关技术研究方面的起步较晚，虽然也取得了不错的研究成果，但仍与国际领先水平有一定的差距。如何结合抗恶劣环境计算机日益增长的需求，继续展开深入研究，进一步提升我国计算机的恶劣环境适应能力，将是国内学者面临的重大机遇和挑战。

参考文献

- [1] L Alkalai, A T Tai. Long-life deep-space applications[J]. IEEE Computer, 1998, 31(4) : 37-38.
- [2] Ann T Tai, Leon Alkalai. On-board Maintenance for Long-Life Systems[C]. //Proc. of ASSET-98. Dallas, USA , 1998 ;69-74.
- [3] 樊林波, 吴映程, 赵明. 软件可靠性与安全性的区别分析及其证明[J]. 计算机科学, 2008, 35(9) : 285-288.
- [4] Markus Hertzman, Rickard Nilsson. A Survey and Evaluation of Diagnostic Tools[D]. Linkoping, 2008.
- [5] Peter Struss, Chris Price. Model based Systems in the Automotive Industry[J]. AI Magazine, 2003, 24(4).
- [6] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, Surya N. Kavuri. A Review of Process Fault Detection and Diagnosis Part I: Quantitative Models-based Methods[J]. Computers and Chemical Engineering, 2002, 27(2003) : 293-311.
- [7] Computers and Chemical Engineering[J]. Strategies, 2003(27) : 313-326.
- [8] IEEE Std 1232 TM—2002 IEEE Standard for Artificial Intelligence Exchange and Service Tie to All Test Environments(AIESTATE)[S]. Piscataway, NJ: IEEE Standards Press.
- [9] J W Sheppard, W R Simpson. A Mathematical Model for Integrated Diagnostics[J]. IEEE Design & Test of Computers, 1991(12) : 25-38.
- [10] XU Shi-yi. Compact-Parity Testing and Testable Design[J]. Journal of Donghua University, 2005(3).
- [11] Meng Jai Tasi, Mango C, T Chao, Jing-Yang Jou, Meng-Chen Wu. Multiple Fault Diagnosis Using Faulty

- Region Identification[D]. 27th IEEE VLSI Test Symposium, 2009.
- [12] Manoj Kumar, A K Verma, A Srividya. Probabilistic Modeling of Network- induced Delays in Networked Control Systems[J].
- [13] Manoj Kumar, A K Verma, A Srividya. Probabilistic Modeling of Networkinduced Delays in Networked Control Systems[J]. International Journal of Applied Mathematics and Computer Sciences, 2009.
- [14] Zahid H Oureshi, Vinay B Sriram. An Integrated Modeling Framework for Military Avionics Mission System Upgrades[J]. IEEE, 2009.
- [15] Tim Holman. Radiation Effects on Microelectronics Short Course[R]. Vanderbilt University Dept. of EECS, 2001.
- [16] K A LaBel, P W Marshall, C J Dale, C M Crabtree, E G Stassinopoulos, M M Gates. SEDS MIL- STD- 1773 Fiber Optic Data Bus: Proton Irradiation Test Results and Spaceflight SEU Data[J]. IEEE Trans. Nucl. Sci, 1993, 40:1638-1644.
- [17] Tanay Karnik, Peter Hazucha, Jagdish Patel. Characterization of Soft Errors Caused by Single Event Upsets in CMOS Processes[J]. IEEE Transactions on Dependable and Secure Computing, 2004, 1(2).
- [18] <https://en.wikipedia.org/wiki/Thomson-CSF>.
- [19] Internet of Things in 2020: Roadmap for the Future[R]. EPOSS.
- [20] Mulligan G. The Internet of Things: Here Now and Coming Soon[J]. IEEE Internet Computing, 2010, 14 (1) :35-36.
- [21] Medaglia C M, Serbanati A. An Overview of Privacy and Security Issues in the Internet of Things[C]. Proceeding of the 20th Tyrrhenian Workshop on Digital Communications, Sardinia, Italy: 2010: 389-395.
- [22] Leusse P, Periorellis P, Dimitrakos T. Self Managed Security Cell, a Security Model for the Internet of Things and Services[C]. Proceedings of the 1st International Conference on Advances in Future Internet, Athens Glyfada, Greece: IEEE, 2009:47-52.
- [23] Hamad F, Smalov L, James A. Energy -aw are Security in M-commerce and the Internet of Things[J]. IETE Technical Review, 2009, 26(5) : 357-362.
- [24] Chen Xiangqian, Makki K, Yen K, et al. Sensor Network Security: A Survey[J]. IEEE Communication Surveys &Tutorials, 2009, 11(2) : 52-73.
- [25] Welbourne, Battle L, Cole G. Building the Internet of Things using RFID: the RFID Ecosystem Experience [J]. IEEE Internet Computing, 2009, 13(3) : 48-55.
- [26] Mealling M. Auto-ID Object Name Service(ONS)v1.0[S]. Auto-ID Center Working Draft, 2003.
- [27] Krylov V, Logvinov A, Ponomarev D. EPC Object Code Mapping Service Software Architecture: Web Approach[EB/OL]. <http://ebajic.free.fr/RFID%20Forum/Papers%20submitted%20but%20not%20presented/EPC%20Object%20Code%20Mapping%20Service.pdf>, 2008.
- [28] Srivastava L. Pervasive, Ambient. Ubiquitous: The Magic of Radin[A]. European Commission Conference From RFID to the Internet of Things[C]. Bmxelles, Belgium, 2006.
- [29] Floerkememr C, Bhattacharyz R, Sarma S. Beyond RFID [C]. Proceedings of TIWDC 2009. Pula, Italy, 2009.
- [30] Sung J, Lopez T S, Kem D. TLI, eEPC Sensor Network for RFID and WSN Integration Infrastructure[C]. Proceedings of IEEE Per-Com W'0771C1. White Plains, NY, USA, 2007: 618-621.
- [31] Botterman M. The European Commission Information Society and Media Directorate General Networked

- Enterprise&RFID Unit-D4[R]. Internet of Things: An Early Reality of the Future Internet Report of the Interact of Things Workshop. Prague, Czech Republic, 2009.
- [32] Buckley J. From RFID to the Interact of Things: Final Report[C]. European Commission Conference From RFID to the Internet of Things. Brussels, Belgium, 2006.
- [33] Toscano M. Department of Defense Joint Robotics Program[C]. AeroSense 2000. Orlando, FL, US: SPIE, 2000: 192-200.
- [34] Spofford J R, Rimey R D. Description of the UGV/Demo II system[C]. Proceedings of Association for Unmanned Vehicle Systems International Conference. US: AUVSI, 1997: 255-264.
- [35] Shoemaker C M, Bornstein J A. Overview of the Demo 111 UGV Program[C]. Robotic and Semi—Robotic Ground Vehicle Technology. US: SPIE, 1998: 202-211.
- [36] Krotkov E, Blitch J. The Defense Advanced Research Projects Agency (DARPA) Tactical Mobile Robotics Program[J]. The International Journal of Robotics Research, 1999, 18(7) : 769-776.
- [37] Fish S. Overview of UGCV and Percept OR Status [C]. Unmanned Ground Vehicle Technology V. Orlando, Florida: SPIE, 2003: 336-339.
- [38] Van Fosson M H, Fish S. Role of Robotics in Ground Combat of the Future: UGCV, Precept OR, and FCS[C]. Unmanned Ground Vehicle Technology 111. Orlando, Florida: SPIE, 2001: 323-327.
- [39] Fish S. UGVs in Future Combat Systems[C]. Defense and Security Symposium. Orlando, Florida: SPIE, 2004: 288-291.
- [40] Jackel L D, Krotkov E, Perschbacher M, et al. The DARPA LAGR Program: Goals, Challenges, Methodology, and Phase I Resuhs[J]. Journal of Field Robotics, 2006, 23(11/12) : 945-973.
- [41] Stentz A, Bares J, Pilarski T, et al. The Crusher System for Autonomous Navigation [C]. AUVSIs Unmanned Systems North America. Las Vegas: Association for Unmanned Vehicle Systems International-Unmanned Systems North America Conference, 2007: 972-986.
- [42] Guo H D. Neural Network Aided Kalman Filtering for Integrated GPS/INS Navigation System [J]. Telkommika Indonesian Journal of Electrical Engineering, 2013, 11(3) : 1221-1226.
- [43] Qian H, An D, Xia Q. IMM-UKF based Land—Vehicle Navigation with Low Cost GPS/1NS[C]. IEEE International Conference on Information and Automation(ICIA). US: IEEE, 2010: 2031-2035.
- [44] Urmson C, Anhah J, Bagne H D, et al. Autonomous Driving in Urban Environments: Boss and the Urban Challenge[J]. Journal of Field Robotics, 2008, 25(8) : 425-466.
- [45] Montemerlo M, Becker J, Bhat S, et al. Junior: the Stanford Entry in the Urban Challenge[J]. Journal of Field Robotics, 2008, 25(9) : 569-597.
- [46] Leonard J, How J, Teller S, et al. A Perception Driven Autonomous Urban Vehicle[J]. Journal of Field Robotics, 2008, 25(10) : 727-774.
- [47] Zuo J, Hoover R, Phillips F. Advanced Thermal Architecture for Cooling of High Power Electronics[J]. IEEE Transactions on Components and Packaging Technologies, 2001, 25(4) : 629-634.
- [48] <http://www.thermacore.com>.
- [49] Ioan S, Greg C, Ravi M, Michele S. Air- coolingExtension- performance Limits for Processor Cooling Applications[C]. Annual IEEE Semiconductor Thermal Measurement and Management Symposium, 2003: 74-81.
- [50] Peterson G P. AnIntroduction to Heat Pipes- modeling, Testing and Applications [M]. John Wiley and

- Sons, New-York, 1994.
- [51] Cotter T P. Principles and Prospects for Micro Heat Pipes [C]. Proc. 5th Int. Heat Pipe Conf., Tsukuba, Japan, 1984, 4: 328-334.
- [52] Itoh A, Polasek F. Development and Application of Micro Heat Pipes [C]. Proc. 7th International Heat Pipe Conference, Minsk, Belarus (CIS), 1990: 295-310.
- [53] Sung Jin Kim, Joung Ki Seo, Kyu Hyung Do. Analytical and Experimental Investigation on the Operational Characteristics and the Thermal Optimization of a Miniature Heat Pipe with a Grooved Wick Structure [J]. International Journal of Heat and Mass Transfer, 2003, 46: 2051-2063.
- [54] 牛文生. 嵌入式计算机抗恶劣环境技术研究综述 [C]. 全国抗恶劣环境计算机第二十三届学术年会, 2013.
- [55] Patz B J, Papelis Y, Pillar R, et al. A Practical Approach to Robotic Design for the DARPA Urban Challenge [J]. Journal of Field Robotics, 2008, 25(8): 528-566.
- [56] 马林立. 外军网电空间战现状与发展 [M]. 北京: 国防工业出版社, 2012.
- [57] 《飞行器控制学报》编辑部. 美军网络司令部正式启动 [J]. 飞行器控制学报, 2010(3): 32-33.
- [58] 邓志宏, 老松杨. 赛博空间概念框架及赛博空间作战机理研究 [J]. 军事运筹与系统工程, 2013, 27(3): 28-31.
- [59] 郝叶力. 研究战场网络不同机理把握“网电一体战”本质内涵 [J]. 军事学术, 2012(5): 33-36.
- [60] <http://www.bjkw.gov.cn/n8785584/2015/src/14.html>.
- [61] 尚利宏, 周密, 胡瑜. 一种基于二阶近似域划分的可重构容错片上系统 [C]. 第六届中国测试学术会议. 2010: 251-257.
- [62] 刑维艳, 王宝军, 刘东. 用 AT91RM9200 构建高可靠嵌入式系统 [J]. 单片机与嵌入式系统应用, 2007(2).
- [63] GJB 1391—1992 故障模式、影响及危害性分析的要求和程序 [S]. 中国国家军用标准, 1992(5).
- [64] 苏永定, 刘冠军, 等. 系统测试性指标确定方法 [J]. 测试技术学报, 2008(5).
- [65] 刘金琨. 智能控制 [M]. 北京: 电子工业出版社, 2005.
- [66] 王宝龙, 等. 基于故障树分析的复杂电子装备诊断策略优化实现 [J]. 弹箭与制导学报, 2006(2).
- [67] 许斌, 周鸣岐. 测试点优化及故障诊断树生成技术 [J]. 国外电子测量技术, 2006(3).
- [68] 张大方, 江招生. 系统级故障诊断的一种优化设计方法 [J]. 湖南大学学报, 1998(6).
- [69] 张大方, 江招生. 基于集团的系统级故障诊断研究 [J]. 计算机学报, 1998, 25(4): 308-314.
- [70] 马羽, 叶琳琳. 1750A 微处理器指令系统浅析 [J]. 科技论坛.
- [71] 高见头, 杨波. CMOS SOI SRAM 电路的抗单粒子能力研究 [J]. 全国抗辐射电子学与电磁脉冲学术年会, 2009.
- [72] 徐尉, 孙力娟, 等. 基于物联网/传感网的智能节能系统 [J]. 计算机研究与发展, 2010, 47: 366-371.
- [73] 吴晓明, 刘祥志, 李刚. 基于物联网技术的数据中心动态节能研究 [J]. 信息技术与标准化, 2014.
- [74] 张浩杰. 不确定环境下基于启发式搜索的智能车辆路径规划研究 [D]. 北京: 北京理工大学出版社, 2012.
- [75] 梁猛, 韩跃, 乔正. 美国《国防部网电空间行动战略》述评 [J]. 国防科技, 2012(2): 25-29.
- [76] 戴清民. 网电一体战引论 [M]. 北京: 解放军出版社, 2002.
- [77] 邓兵, 刘峰. 未来信息作战的主要形式——“网电一体战” [J]. 航天电子对抗, 2002(1): 47-49.
- [78] 侯印鸣, 李德成, 孔宪正, 陈素菊. 综合电子战 [M]. 北京: 国防工业出版社, 2000.
- [79] Okojie R S, Savrun E, Nguyen P, et al. Reliability Evaluation of Direct Chip Attached Silicon Carbide Pressure Transdacers [A]. 3rd International Conference on Sensors [C]. Vienna, Austria 2004: 22-30.

作者简介

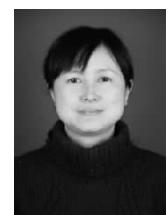
沈 崩 中国航天科工集团第二研究院七〇六所，研究员，主要研究方向为计算机体系结构、加固技术、信息安全技术等，shensong711@163. com。



孙丽婷 中国航天科工集团第二研究院七〇六所，博士，主要研究方向为环境感知技术、网电技术等，sunliting@163. com。



杨 帆 中国航天科工集团第二研究院七〇六所，研究员，主要研究方向为计算机体系结构、加固技术等，yang_fan_liu@163. com。



祁春慧 中国航天科工集团第二研究院七〇六所，博士，主要研究方向为无线通信技术、信息安全技术等，qichunhui706@163. com。



关键词索引

- NewSQL 144,181,184,185
传输 1,2,3,7,8,10,11,12,13,14,15,18,19,20,21,22,23,24,26,27,28,29,30,31,32,36,48,66,76,77,102,112,114,121,146,153,160,162,176,177,178,328,329,332,333,339,346,347,366,369
大数据 13,32,38,39,41,42,43,44,50,51,52,53,54,55,59,61,62,64,67,100,104,114,121,122,123,124,125,127,129,130,132,133,134,142,143,144,145,146,147,148,149,150,158,164,165,166,172,173,174,175,178,179,180,188,189,190,191,200,201,204,236,237,253,254,297,328
大数据机器学习 121,127,129,130,133,134
大数据学习 121,122,123,124
多模态 131,132,189,282,284,285,286,287,290,292,314
恶劣环境 326,330,361,362,363,364,365,366,367,368,369,370,371,372,374,376,377,378,379,382
分布式学习 121,122,126,127,128,132,133
服务发现 253,254,255,256,261,262,263,266,267,268,270,279
服务推荐 253,256,261,263,264,267
服务组合 253,254,256,257,258,261,262,263,264,266,267,268,269,270,278
工业控制计算机 324,325,326,327,329,330,334,335,336,337,338,339,343,350,351,357,358,359,360
故障注入 66,69,71,72,77,78,79,80,81,87,91,93,364,371,372,376,379
哈希学习 121,122,123,124,125,131,132,133,134,142,290
互联网+ 38,39,41,42,47,48,49,50,51,52,53,54,55,57,59,61,61,62,120,178,181,253
互联网交通 38,51,55,62
互联网金融 38,39,47,50,51,52,57,61,106,113,114,116,117,118,120
环境感知技术 361,362,363,366,367,374,375,376,378,379,383
机器学习 55,121,122,124,125,126,127,128,129,130,133,134,142,145,147,180,189,205,208,210,211,213,214,215,232,233,234,235,236,237,254,262,283,293,296,301,312,315
基础计算系统 66,67,68,69,75,77,78,86,87,91
加固防护技术 361,362,363,368,375,376,377,378,379
经验软件工程 203,204,205,206,209,210,215,218,219,223,225,227,228,229,235,236,237,251,252
科技与应用 97
可靠性 66,69,70,71,72,73,75,76,77,78,81,82,83,84,85,86,87,88,89,90,91,93,94,95,96,118,155,162,171,184,202,209,214,325,326,327,328,329,330,333,336,337,338,340,341,343,349,350,351,352,354,355,356,357,358,361,362,363,364,371,376,377,378,379
流处理 144,145,146,147,148,159,161,162,163,164,165,197
路由 1,13,14,15,17,22,24,25,27,28,29,30,32,35,36,47,70,76,77,159,160,163,265,343,344,369
媒体计算 282,283,284,292,297
嵌入式系统 46,214,324,325,326,330,339,340,343,349,350,356,382
容错计算机 93,324,325,350,351,352,353,355,357,360,363

- 容错体系结构 361
软件工程 81, 82, 119, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 215, 216, 218, 219, 223, 224, 225, 227, 228, 229, 234, 235, 236, 237, 251, 252, 280, 281, 371
软件可靠性 66, 69, 72, 73, 77, 81, 82, 87, 87, 91, 93, 94, 96, 354, 357, 361, 362, 363, 364, 371, 376, 377, 379
深度学习 122, 133, 134, 135, 282, 283, 284, 285, 286, 287, 289, 290, 291, 292, 293, 301, 305, 314
视频分析 56, 282, 288, 289, 290
视频检索 282
数据到文本的生成 298, 299, 308, 309, 310, 311, 312, 316
数据分析 41, 43, 44, 53, 54, 55, 121, 147, 148, 150, 156, 157, 158, 171, 173, 174, 175, 176, 177, 178, 179, 180, 181, 189, 190, 191, 199, 200, 201, 203, 205, 209, 236, 309
数据管理系统 144, 145, 148, 149, 151, 153, 157, 158, 160, 161, 163, 165, 166, 171, 181, 189, 191, 193, 197
随机投影 121, 122, 125, 126, 132, 133
随机优化 121, 122, 128, 129, 132, 134
体系结构 1, 2, 13, 20, 21, 22, 23, 24, 25, 33, 35, 55, 57, 82, 87, 93, 119, 147, 150, 189, 218, 219, 220, 221, 235, 245, 252, 264, 324, 334, 341, 343, 344, 345, 346, 349, 351, 353, 355, 357, 360, 361, 362, 363, 365, 366, 370, 371, 376, 377, 379, 383
天地一体化网络 1, 2, 3, 8, 19, 20, 24, 25, 26, 27, 28, 31, 32, 33, 35, 36, 37
图数据 144, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 172, 179, 188, 191, 200, 367
图像到文本的生成 298, 299, 312, 315, 316
网络可靠性 65, 66, 69, 75, 76, 77, 82, 83, 84, 85, 86, 87, 88, 90, 91, 93, 94, 95
网络信息服务 2, 38, 48, 49, 55, 64
网络与信息安全 98, 100, 114
维修策略 66
文本到文本的生成 298, 299, 305, 316
业务流程管理 253, 254, 258, 259, 260, 261, 265, 267, 268, 269, 270
意义到文本的生成 298, 299, 305, 306, 307, 308, 309, 316
云计算 38, 39, 44, 45, 46, 47, 48, 49, 50, 52, 55, 63, 66, 67, 68, 69, 71, 72, 73, 77, 78, 81, 87, 88, 89, 91, 93, 99, 112, 114, 116, 119, 120, 122, 165, 185, 186, 188, 251, 253, 268, 269, 280, 297, 340
在线分析 56, 144, 174, 175, 179
知识型服务计算 253, 254, 255, 270
众包 50, 122, 144, 148, 165, 166, 167, 168, 169, 170, 171, 172, 173, 191, 197, 198, 233, 234, 235, 293, 305
自然语言生成 298, 299, 300, 301, 303, 308, 309, 310, 314
综述 1, 38, 63, 65, 66, 69, 70, 91, 93, 98, 158, 197, 206, 207, 219, 221, 257, 263, 270, 301, 382
组网结构 1, 2, 27

作者索引

- 陈 刚 192, 202
从立钢 37
崔 烽 192, 201
底晓强 37
丁志军 65
冯岩松 316, 323
高 军 192, 201
龚 健 360
郝文江 121
何克清 279, 280
江 卓 36
姜会林 37
蒋昌俊 64, 65
李翠翠 121
李飞飞 192, 202
李国良 192, 201
李贺武 35
李武军 133, 143
梁 鹏 221, 252
刘宏伟 92, 94, 95, 97
刘显著 37
刘 鑫 335, 336, 358, 359, 360
陆 洲 36
孟小峰 192, 200, 201
莫毓昌 95, 96
祁春慧 383
祁 晖 37
秦智超 36
沈 桢 383
孙丽婷 383
孙薇薇 316, 323
万小军 316, 322
王俊丽 64, 65
王鹏伟 65
王 青 217, 229, 234, 235, 251
王秋月 192, 202
王尚广 279, 280
王天枢 37
王忠杰 281
吴 飞 133, 296
吴建平 33, 35
吴 健 278, 281
吴 茜 35
徐丽萍 121
徐明伟 35
杨 帆 383
杨 桦 335, 360
杨孝宗 79, 92, 96, 355
杨增印 36
于俊清 297
禹晓辉 192, 198, 201
张 贺 251
张利军 133, 143
张 莉 252, 279
张 平 36
张 昭 96
钟发荣 95, 97
周明辉 252
周志华 143, 217
朱文武 133, 297
左德承 94, 96