

中国计算机学会通讯



COMMUNICATIONS OF THE CCF

第15卷 第2期

总第156期 2019年2月



大数据共享与交易 P8

量子计算五人谈 P46

专访CCF杰出贡献奖得主戴维·阿兰·格里尔 P66



ADL 学科前沿讲习班

The CCF Advanced Disciplines Lectures

站在学科前沿 打开技术之门

高水平 (资深专家讲授)

立前沿 (最新和热点技术)

大剂量 (三整天)

ADL 是 CCF 举办的学科前沿讲习班，目的是使青年学者短期内深入了解计算领域某个学科前沿发展动态，开拓眼界，为尔后的科学研究打下基础。2009 年开始举办，每年 10 期。

联系：adl@ccf.org.cn 188 1066 9757



中国计算机学会通讯 COMMUNICATIONS OF THE CCF



主办 中国计算机学会
China Computer Federation

刊名题字 张效祥

编辑 《中国计算机学会通讯》编辑部
编辑部主任：李梅
地址：北京市海淀区科学院南路6号
通信：北京 2704 信箱 100190
电话：(010) 6267 0365
传真：(010) 6252 7485
http://www.ccf.org.cn
E-mail: cccf@ccf.org.cn
封面设计：SEEKLAB

声明

《中国计算机学会通讯》(CCCF)刊登的文章，除 CCF 或 CCCF 特别署名外，仅代表作者的学术观点。CCCF 鼓励与支持学术争鸣。

版权声明

中国计算机学会 (CCF) 拥有《中国计算机学会通讯》所刊登内容的所有版权，未经 CCF 允许，转载本刊文字及照片会被视为侵权，CCF 将追究其法律责任。

编辑单位：中国计算机学会
印刷单位：北京华联印刷有限公司
发送对象：中国计算机学会会员
印刷日期：2019 年 2 月

主编

李国杰 CCF 名誉理事长，CCF 会士，中国工程院院士

执行主编

钱德沛 CCF 会士，北京航空航天大学教授，中山大学计算机学院院长

专题

主编 袁晓如 CCF 理事，北京大学研究员
编委 陈熙霖 CCF 会士、理事，中国科学院计算技术研究所研究员
李向阳 CCF 专业会员，中国科技大学教授
廖小飞 CCF 高级会员，华中科技大学教授
王蕴红 CCF 会士、理事，北京航空航天大学教授
杨珉 CCF 专业会员，复旦大学教授
郑宇 CCF 杰出会员，京东集团副总裁

专栏

主编 彭思龙 CCF 理事，中国科学院自动化研究所研究员
编委 包云岗 CCF 理事，中国科学院计算技术研究所研究员
郭得科 CCF 杰出会员，国防科技大学教授
徐恪 CCF 理事，清华大学教授
王涛 CCF 理事，爱奇艺公司首席科学家
王长虎 CCF 高级会员，字节跳动人工智能实验室总监

动态

主编 唐杰 CCF 杰出会员，清华大学教授
编委 鲍捷 CCF 专业会员，北京文因互联科技有限公司 CEO
黄萱菁 CCF 高级会员，复旦大学教授
蒋洪波 CCF 杰出会员，湖南大学教授
刘知远 CCF 高级会员，清华大学副教授
宋国杰 CCF 高级会员，北京大学副教授
俞扬 CCF 专业会员，南京大学副教授

译文

主编 卜佳俊 CCF 常务理事，浙江大学教授
编委 胡春明 CCF 理事，北京航空航天大学副教授
姜波 CCF 理事，浙江工商大学教授
苗启广 CCF 理事，西安电子科技大学教授

学会论坛

主编 杜子德 CCF 秘书长
编委 胡事民 CCF 会士、常务理事，清华大学教授

CONTENTS 目录

2019年2月 第15卷 第2期 总第156期



2018CCF 颁奖大会摄影：明理

大数据共享与交易

数据共享是一个已得到较大关注的问题，而数据交易则不然。在实践中我们经常看到，在数据共享管控方面，要么失之过严，使应该共享的数据得不到使用，造成数据资源的浪费；要么失之过宽，导致个人数据的滥用，个人隐私被侵犯。如何把数据当作一种资产，确定其权属，评定其价格，并像货币那样通过“数据银行”等手段而流通使用，在实现数据价值的基础上实现其共享，的确是一个新的命题。当然，要真正实现数据的交易，不仅需要必要的技术机制和手段的支持，更需要国家在法律、政策层面的作为，希望本专题能对此有所推动。

(钱德沛 P8~45)

卷首语

7 科学、技术和工程

杜子德

专题

8 大数据共享与交易

特邀编辑：陆品燕 吴帆

11 移动社交网络大数据下的营销计算

张远行 边凯归 宋令阳 等

17 盘活大数据资产——个人大数据权益保护与合理使用

段旭良 郭兵 吴帆 等

23 安全多方计算与数据流动

安瑞 谢翔 孙立林

30 数据共享和交易——数据的质量、价值和价格

张兰 李向阳 李安然 等



阅读整本

敬告读者

欢迎读者提出意见或建议。

编辑部联系方式：

电话：(010)6267 0365

E-mail: cccf@ccf.org.cn

查阅电子版：

<http://dl.ccf.org.cn/cccf/list>



个人数据隐私保护与流通使用的平衡
(详见段旭良专题文章)

36 数据交换

左 淼

39 个人数据交易：从保护到定价

牛超越 郑臻哲 吴 帆 等

专栏

46 量子计算五人谈

孙贤和

52 区块链的极限

特邀专栏作家：万 赞

58 机器阅读理解：如何让计算机读懂文章

朱晨光

64 The CS David 专栏

动物计算

作者：戴维·阿兰·格里尔 (David Alan Grier)

特邀译者：孙晓明

动态

66 用“文章”架起交流的桥梁——专访 CCF 杰出贡献奖得主戴维·阿兰·格里尔

韩玉琦

70 2018 自然语言处理实证方法会议概览

岂凡超 韩 旭 刘知远

74 新技术 & 新应用

译文

76 机器学习如何影响本科生计算机课程

作者：本杰明·夏皮罗 (R. Benjamin Shapiro)
丽贝卡·菲布林克 (Rebecca Fiebrink)
彼得·诺维格 (Peter Norvig)

译者：刘如意 史媛媛 苗启广

学会论坛

82 2018 CCF 颁奖大会在京举行

94 读编往来

信息索引

• CCF ADL	封二
• CCF CSP 第 16 次认证	6
• 2018 NOI 教师培训圆满收官	29
• CCF 表彰 2018 年度优秀专业委员会	57
• CCF 专委发展 & 交流会在京举行	63
• CCF 教育工委走进高校组召开工作会议	63
• CCF 秘书处总结及表彰年度优秀	72
• CCF 认定第二批“CCF 中国计算机历史记忆”	85
• CCF 将举行理事会换届选举	86
• CCF 常务理事会最后一轮执行委员产生	87
• CCF 生物信息学专业组升级为专业委员会	87
• 七名 CCF 杰出会员当选 CCF 会士	88
• CCF 走进高校	89
• CCF 表彰 2018 年度优秀会员活动中心	90
• CCF 表彰“会员发展优秀奖”	90
• 98 位 CCF 专业会员晋升为高级会员	91
• 26 位 CCF 高级会员晋升为杰出会员	92
• CCF 增设五个学生分会	92
• 49 名讲师被评为 2018 年度 CCF 杰出演讲者	93
• CCF 会员活动中心动态 (2018 年)	96
• CCF 会员续费	封三
• CCF 颁奖大会	封底

Preface

7 Science, Technology and Engineering

Du Zide

Features

8 Data Sharing and Trading

Guest Editors: Lu Pinyan and Wu Fan

11 Marketing Computing under the Big Data of Mobile Social Networks

Zhang Yuanxing, Bian Kaigui, Song Lingyang and Li Xiaoming

Marketing computing is an important means of marketing activities in mobile social networks. This article analyzes the influence of social events on mobile social network marketing based on the information dissemination data from WeChat, and introduces the state-of-the-art approaches and the objectives of marketing computing. Finally, the research challenges and opportunities of marketing computing are summarized.

17 Revitalizing Big Data Assets: Rational Use and Protection of Personal Big Data

Duan Xuliang, Guo Bing, Wu Fan and et al.

In the era of digital economy, personal data is not only a personal asset, but also an important resource for economic and social development. Personal data bank is a new mode of personal big data asset management and value-added service based on the mature monetary asset management architecture. Under the premise of protecting the ownership, privacy, usufruct and the right to know of personal data, the new data bank mode makes the whole process of data sharing, circulation and usage transparent and controllable. The healthy development of personal data circulation market will greatly help to revitalize the valuable asset of personal big data.

23 Secure Multi-Party Computation and Data Flow

An Rui, Xie Xiang and Sun Lilin

Data has become the most valuable resource in the world. However, there are so many data privacy security problems and isolated data islands at present. As a collaboration computing technology, Secure Multi-Party Computation (MPC) can provide a strong fulcrum for realizing data flow under the premise of guaranteeing the data privacy and security of all participants. This article explains the principle of an MPC framework in detail. Then it analyzes the MPC technology from the view of protocol principle and describes its engineering achievement briefly.

30 Data Sharing and Transaction: Data Quality, Value and Price

Zhang Lan, Li Xiangyang, Li Anran and Xue Shuangshuang

As data becomes valuable increasingly, data trading and sharing markets are booming. This article reviews related research results, proposes several primitive solutions and identifies some challenging open problems for three crucial issues of data trading: the quality of data, the value of data, and the price of data.

36 Data Exchange

Zuo Song

This article focuses on the differences of valuation models between the digital goods and the normal goods, which come from some special properties of the former, and how the differences influence the strategic behavior of the agents who participated in the data exchange market. The differences mainly come from the zero copy-cost property of data and the negative externality from sharing data with potential competitors. Because of these properties, both the strategies and the Nash equilibrium structure in data exchange markets become much more complicated than those in normal goods exchange markets.

39 How to Trade Personal Data? From Protection to Pricing

Niu Chaoyue, Zheng Zhenzhe, Wu Fan and Chen Guihai

This article investigates personal data trading in data markets from the standpoint of data broker, and covers three major topics: verifiable and privacy preserving data sharing ecosystem, rigorous quantification and compensation of privacy losses to data contributors, and robust pricing of noisy data services for data consumers. On each topic, the related work, existing problems and challenges, preliminary designs, and open questions are proposed.

Columns

46 A Five Men's Conversation on Quantum Computing

Xian-He Sun

Quantum computing is a subject of rapid development as well as a subject of great interest. During the New Year holiday, some scholars in the United States had a lively discussion on quantum computing online and offline. We invited professor Xian-He Sun to edit their conversation into an article and publish it here, for our readers.

52 The Limit of Blockchain

Wan Yun

The blockchain technology was originally invented to time-stamp electronic documents and provide a distributed form of authority and credibility compared with the traditional centralized form. The popularity of bitcoin increased its visibility as well as people's unrealistic expectation. Blockchain will not replace the Internet or the Web due to its restrictive built-in data organizing infrastructure. Blockchain is useful in various scenarios, such as digital currency or supply chain management. However, its usefulness needs to justify the relatively high maintenance cost.

58 Machine Reading Comprehension: How to Teach Computers to Read Articles

Zhu Chenguang

With the rapid development in AI, we have made tremendous improvement in natural language processing (NLP). In recent years, one of the core tasks

in NLP, machine reading comprehension, has been under intensive study by researchers. Machine reading comprehension (MRC) is a technique to help computers understand articles and answer related questions. Machine reading comprehension has a variety of applications in information industry. In this article, we introduce the background, models, and challenges of MRC.

64 The CS David

Computing with Animals

David Alan Grier (translated by Sun Xiaoming)

This article first introduces a conference on Animal-Computer-Interaction and its three major themes. Although it is far from clear that Animal-Computer-Interaction will have the same kind of impact as the original interplay of computing and agriculture, by studying how animals interact with machines, we may get a better understanding of intelligence and learn a little more about what intelligence is not and how hard it is to capture in computing systems since animals are supposedly simpler creatures.

Advances

66 Building a Communication Bridge with “Articles”: An Interview with CCF Outstanding Contribution Award Winner David Alan Grier

Han Yuqi

70 An Overview of the EMNLP2018

Qi Fanchao, Han Xu and Liu Zhiyuan

74 New Technologies & New Applications

Translations

76 How Machine Learning Impacts the Undergraduate Computing Curriculum

Shapiro R B, Fiebrink R and Norvig P (translated by Liu Ruyi, Shi Yuanyuan and Miao Qiguang)

The growing importance of machine learning creates challenging questions for computing education. In this article, we consider how machine learning might change what we consider to be core CS knowledge and skills, and how this should impact the design of both machine learning courses and the broader CS university curriculum.



CSP

让业界认可你的
专业能力

CCF CSP 软件能力认证 CERTIFIED SOFTWARE PROFESSIONAL

第16次认证 2019.3.17



考查算法设计和编程能力
名校名企入门参考凭证



CCF统一命题，统一评测
5道大题，同一时间测试

全国各大城市100余所高校

报 名 : cspro.org
咨 询 : csp@ccf.org.cn
电 话 : (010) 6260 0321-16



合作高校及企业：(排名不分先后)



清华大学



北京航空航天大学



国防科技大学



北京大学



上海交通大学



中国农业大学



华中科技大学



哈尔滨工业大学



西安交通大学



电子科技大学



天津大学



山东大学



中山大学



湖南大学



南京理工大学



HUAWEI



Bai度一下 你就知道



Tencent

腾讯



阿里巴巴

Alibaba.com



360

金蝶, 企业管理专家



Kingdee

金蝶, 企业管理专家



Microsoft



Intel



滴滴



中国优游彩票网



卷首语

CCCF 2019 年第 2 期

CCF 秘书长 杜子德

科学、技术和工程

中国常把科学和技术放在一起说成“科技”，也有“科学技术是第一生产力”的说法。这种含混的说法导致评价和行为上的错误，对中国的科学和技术发展有害。

科学 (science) 的作用是发现和解释自然或社会现象，爱因斯坦的相对论、生物学中的 DNA 双螺旋结构模型等是自然科学的例子，荣格的集体无意识是社会科学的例子。科学能发现规律，并解释规律让人类了解自然，但并不能直接创造经济价值，所以科学还不是生产力。“科学技术是第一生产力”是指技术，而不是科学。

技术 (technology) 是人类在长期利用和改造自然的过程中积累起来的知识、经验、技巧和手段，近代以来发明的技术大多基于科学理论。比如知道了激光原理以后，发明了激光通信、激光加工等技术，广泛用于各个领域。但技术发明并不总是依赖科学原理，有时在不懂科学原理前也会有技术发明，但难以走远，比如中国发明了火药，但并不知道其化学原理，因而没有发明火炮。

工程 (engineering) 是用现有成熟的技术（群）和其他元素完成一项有实用价值的人造物或产品，曼哈顿计划、两弹一星、北斗系统等都是工程。很好的技术并不意味着很好的工程，只有将技术成功运用在工程中，才能取得预想的效果。

科学要回答“是什么”和“为什么”的问题，技术则回答“做什么”和“怎么做”的问题。没有科学原理，就没有先进技术的发明，没有足够多的实用技术，就不可能完成工程任务。科学、技术和工程相互之间有密切的关系，但均有各自的特点和规律。因此，从事科学、技术研究和工程建设必须遵从不同的方法论，也有不同的评价方式。

科学是探索性的，事先并不知道会有什么结果，往往是长期积累，偶然发现。因此，对于从事科学的研究的人，要给予充分的自由度，不能要求什么时候一定要出成果。对于技术发明，尽管有明确的指向性和可达成的目标，但也需要积累，具备一定条件，比如对于原理的理解，对于相关材料的要求以及符合要求的加工工具等。实现工程目标除了要求技术成熟之外，还需要一套复杂而严密的管理手段。

科学、技术和工程有不同的特点，其评价标准相差很大。中国往往把三者混为一谈，把科学看成技术，甚至看成工程，往往采用工程管理的办法对待基础研究和高技术研发。如果原理还没有搞清楚就匆匆上马做大事，无疑是拔苗助长，造成巨大的浪费。本期的《量子计算五人谈》为理解科学、技术和工程解剖了一个鲜活的案例：通用的量子计算和量子通信目前属于基础研究阶段，还不到盲目上大工程的时候。

基础科学研究应主要靠公共财政投入，由科学家自由探索。技术可以创造价值，其投入应主要由市场来决定，而不是政府，但关系到国家安全和公共福利的技术研究，则需要国家投入，如两弹一星。发展 CPU 和操作系统等核心技术的关键是培育产业生态，需要长期的技术积累，最终只能靠有能力的骨干企业来解决。

尽管科学与技术的边界往往难以非常清晰地界定，但不了解科学、技术和工程的本质差异，就会犯极大的错误，现在是正本清源的时候了。

杜子德

大数据共享与交易

特邀编辑：陆品燕¹ 吴帆²

¹ 上海财经大学

² 上海交通大学

关键词：大数据共享 大数据交易

在大数据时代，数据拥有巨大的经济价值，被喻为新兴的石油资源。通过深入挖掘跨领域的数据资源，发现数据背后的经济规律，能有力促进产业升级和跨越式发展。大数据已经成为世界各国关注的焦点，我国也已经把大数据列入国家战略发展方向，并希望联合人工智能一起助力中国的经济腾飞。机器学习算法的突破性进展和人工智能技术的大范围落地应用也离不开海量高质量数据的供应^[1]。各个组织机构都希望获取有价值的数据资源来优化性能和辅助决策。然而当前的数据共享与流通规则和技术却无法满足各类应用对于数据资源的强烈需求，形成了大量与世隔绝的数据孤岛，这是对数据资源的极大浪费。因此，亟须支持数据共享和交易的开放平台和相关技术来打破数据壁垒，连接数据孤岛，促进数据在互联网上的流通，以挖掘大数据的经济价值，释放各类数据的应用潜力。

数据的流通和交易作为新兴的商业模型，已经引起企业界和学术界的高度重视。美国的 Xignite^[2] 公司运营着金融行业的数据共享，Gnip^[3] 公司出售来自社交网络的数据，Sabre^[4] 公司则交易旅行用户的订阅和查询信息。国内数据交易市场也呈现井喷式发展的态势，贵阳大数据交易所^[5] 是国内第一家大数据交易所，随后上海数据交易中心^[6]、武汉东湖大数据交易中心^[7] 等类似的平台也相继出现。近期，基于区块链的分布式数据交易市场更是在业界掀起了一股热潮，如 IOTA IoT 数据市场^[8]、Data-broker Dao^[9] 和 BAIC^[10] 等。在学术界，美国华盛顿

大学 Dan Suciu 教授所领导的研究组是数据交易方向的开拓者，并形成了一系列相关工作^[11,12]。中国科学技术大学李向阳教授团队在《美国计算机学会通讯》(Communications of the ACM, CACM) 专栏^[13] 和《中国计算机学会通讯》^[14] 中分析了国内数据共享与交易市场的机遇以及在数据预处理、数据质量评估、数据定价、数据安全隐私与数据追溯等方面面临的挑战。国家的各项大数据发展文件也频繁出现“大数据交易”相关的关键词。比如在 2015 年国务院印发的《促进大数据发展行动纲要》中明确提出“要引导培育大数据交易市场，开展面向应用的数据交易市场试点，探索开展大数据衍生产品交易，鼓励产业链各环节的市场主体进行数据交换和交易，促进数据资源流通，建立健全数据资源交易机制和定价机制，规范交易行为等一系列健全市场发展机制的思路与举措”。工业和信息化部于 2017 年 1 月发布的《大数据产业发展规划（2016 – 2020 年）》指出“要开展数据资源分类、开放共享、交易、标识、统计、产品评价、数据能力、数据安全等基础通用标准以及工业大数据等重点应用领域相关国家标准的研制。”

不同于传统的商品，数据作为一种非独占性的特殊资源，具有增长速度快、复制成本低、潜在价值未知、所有权难确定、流通渠道难管控等特性，为构建高效、可信、公平、安全的数据共享与交易市场带来了诸多挑战。

异构海量数据预处理：要使数据成为有价值

的资产，我们需要对数据进行适当的预处理，如数据清洗、标准、校准、融合与脱敏。一方面，各领域不同模态的数据具有迥异的数据特性，比如金融数据具有强时间关联性，感知数据具有时空关联性。另一方面，中国人工智能产业的快速增长需要海量高质量的数据，而手工标注数据的方式仍然非常普遍。如何为异构海量数据提供高效的自动化预处理方法是数据交易市场需要解决的首要问题。

数据确权与数据验证：在共享交易数据之前，我们应该明确数据资产的各项权利，包括数据的所有权和使用权。个人日常活动所产生数据的所有权毫无疑问属于个人。然而，数据不同于传统的商品，具有看过即拥有的特性，难以清晰地界定所有权。当前大数据行业大都采取以服务换取数据的方式，混淆数据的所有权和使用权，而数据所有者无从知晓和管控自身数据的使用情况。数据市场需要高效准确的数据溯源方法以达到数据来源可查，使用可查以及流向可查，并积极推进数据权益保护的相关立法。高质量的真实可信数据是数据共享交易平台的基石。数据本质上是二进制符号，卖家可以任意伪造虚假数据，例如通过对抗神经网络生成的图片数据能达到以假乱真的效果。数据交易市场需要可行的验证方法来确保数据来源的真实性、数据质量的可靠性以及数据计算结果的准确性。

数据安全与隐私保护：大量数据中包含着丰富的敏感信息，数据市场的交易安全与隐私保护对于敏感数据的共享显得尤为重要。一方面，个人隐私数据能够用来提供个性化服务与精准营销，具有较高的价值，是数据共享与交易的热门资源。另一方面，由于数据交易市场汇聚有海量的个人数据资源，数据的泄露将造成难以估量的后果，比如剑桥分析公司通过攻击 Facebook 百万用户资料以影响美国的政治选举。因此，数据交易平台有责任和义务对用户数据进行安全存储和传输，在充分保护用户隐私的前提下合理使用数据资源，在开放流通和隐私保护之间找到合适的平衡点。

数据质量和价值评估：为了保障数据交易参与方的权益，构建公平可信的规范市场，维护健康

的数据交易生态，数据的质量评估和价值评估成为亟待解决的难题。质量评估关注数据内容本身的多维度特性，如完整性、准确性、精确性、一致性、时效性等，而价值评估则在评估数据质量的同时进一步综合考虑数据在生产过程中的成本和在不同任务中的产出。相关研究目前还面临许多挑战，比如特征难以量化评估，数据集合质量评估效率低，数据成本难计算，数据使用价值难预测，数据价值动态变化，等等。数据的易复制、流通渠道难管控的特点使得售出数据难以实现退货，这进一步提高了在售前对数据进行准确可信的质量和价值评估的要求。

数据定价与收益分配：数据共享交易市场的持续健康发展需要合理的数据定价机制与公平的收益分配策略。首先，数据作为信息系统的副产品，其成本估计困难。其次，数据的市场价值受到应用场景和市场上同类产品的影响，例如 GPS 数据在导航应用中价值较高，在金融征信应用中则价值较低。数据应用场景的多样化和动态性大大增加了数据的市场价值评估的难度。此外，数据复杂的关联性使得数据市场的套利行为更加普遍：数据买家可以通过低价数据来推断高价数据内容。如何克服以上困难，设计合理的数据定价机制来使数据资源得到更合理的分配，并实现交易参与方的多赢，是数据交易可持续进行的重要问题。

本期专题的六篇文章对以上方向做了初步探索。

在《移动社交网络大数据下的营销计算》一文中，北京大学“博雅”特聘教授宋令阳所在研究团队系统地介绍了如何使用移动社交互联网大数据来进行精准营销计算。作者进一步阐述了如何利用微信生态圈的各类数据重构信息传播网络，使社交网络中的营销影响力达到最大，并提供用户的匹配、识别和推荐等任务。该文章向我们展示了流通的大数据所能催生的各类应用。

四川大学教授郭兵所在研究团队撰写了《盘活大数据资产——个人大数据权益保护与合理使用》。作者重点关注个人大数据的资产化管理、运营和服务，以及保护用户的隐私和数据安全。讨论如何完善个人大数据领域的政策法规，在保障用户合法权

利的前提下，充分挖掘个人数据的价值，促进个人数据流通交易市场的良性发展，盘活个人大数据这一宝贵资产。作者将解决思路整合成一套个人大数据资产化的管理方案——个人数据银行，并通过实践来衡量其产业影响。

在《安全多方计算与数据流动》一文中，矩阵元技术（深圳）有限公司安瑞等人采用安全多方计算的具体技术来保护数据市场各参与方的数据隐私安全，并完成协同计算，为数据流动的真正实现提供安全保障。作者从协议原理的角度对可信多方计算进行了分析，并对其工程化结果进行简要说明，最后阐述了如何将安全多方计算的技术应用于医疗数据和人工智能数据共享的场景中。

中国科学技术大学教授李向阳等人在《数据共享和交易：数据的质量、价值和价格》一文中，围绕数据“质量如何？”“值多少钱？”“卖多少钱？”三个问题对当前数据质量评估、价值评估和数据定价的市场和科研现状进行了简要总结，针对目前存在的部分挑战提出了数据价值全面评估和时间敏感数据定价的初步解决思路，并讨论了相关的开放性问题，为未来的研究提供参考。

谷歌研究院的左淞在文章《数据交换》中，通过分析数据资源的价值、稀缺性及共享性等特殊性，通过博弈论建立了一个数据交换的经济学模型，并分析其市场均衡的存在性和算法复杂度。

在《个人数据交易：从保护到定价》一文中，上海交通大学教授吴帆所在研究团队介绍了个人数据交易市场的架构，并在安全可信数据交易环境的搭建、隐私泄露量化和补偿、隐私数据服务定价机制设计三个主要方面进行了初步的探索。作者详细讨论了在保护数据市场各方敏感信息的基础上实现数据处理结果的可验证性，以达成私密信息保护和数据服务高可用性的统一。为补偿用户在数据交易中的隐私泄露，作者进一步采用差分隐私工具量化用户隐私泄露程度，设计了尊重隐私的干扰型数据服务的定价机制。

以上六篇文章从关系经济民生的多方面数据需求出发，对数据共享交易各个环节中面临的关键科

学挑战进行了剖析，也给出了一系列创新的思路和方法。当然，大数据的共享和交易仍处在起步阶段，这是一个朝气蓬勃、充满机遇与挑战的领域。我们需要更多的力量投入到相关科研难题的研究和实际应用的验证中，从而推动数据共享和流通，为经济发展带来强劲的新动力。 ■



陆品燕

CCF 杰出会员，2014 年 CCF 青年科学家奖得主。上海财经大学教授、理论计算机科学研究中心主任。主要研究方向为计算复杂性、算法、算法博弈论。
lu.pinyan@mail.shufe.edu.cn



吴帆

CCF 专业会员。上海交通大学计算机科学与工程系教授。主要研究方向为网络经济学、无线网络、移动计算、隐私安全。
fwu@cs.sjtu.edu.cn

参考文献

- [1] Sun C, Shrivastava A, Singh S, et al. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era[C]//Proceedings of the 2017 IEEE International Conference on Computer Vision. IEEE CS, 2017: 843-852.
- [2] Xignite[OL]. <http://www.xignite.com/>.
- [3] Gnip[OL]. <https://gnip.com/>.
- [4] Sabre[OL]. <https://www.sabre.com/>.
- [5] 贵阳大数据交易所 [OL].<http://www.gbdex.com/>.
- [6] 上海数据交易中心 [OL].<https://www.chinadep.com/>.
- [7] 武汉东湖大数据交易中心 [OL].<http://www.chinadatatrading.com/>.
- [8] IOTA 区块链数据交易市场 [OL]. <https://data.iota.org/>.
- [9] Databroker Dao[OL].<https://databrokerdao.com/>.
- [10] BAIC[OL].<http://baic.io/>.
- [11] Balazinska M, Howe B, Suciu D. Data markets in the cloud: An opportunity for the database community[J]. *Proceedings of the VLDB Endowment*, 2011, 4(12): 1482-1485.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

移动社交网络大数据下的营销计算

关键词：移动社交网络 大数据 网络嵌入向量 营销计算

张远行 边凯归 宋令阳 等
北京大学

移动社交网络中营销计算的概念及框架

社交网络作为一种集合人脉资源的平台，通过刻画关系网建立市场渠道，满足了用户的精神和物质需求^[1]。同时，移动社交网络应用普遍推出了移动支付的功能，为用户提供了便捷的支付渠道，带来了额外的经济效益，如移动互联网中的内容打赏、信息流广告展示、微商交易及虚拟物品交易等。借助于大数据和移动社交技术，社交应用呈现显著的移动化、本地化特征，是很好的商业导流入口。对一般用户而言，移动社交网络提供了便捷的消遣方式；对企业或个体商户来说，移动社交网络提供了新的推广营销渠道。根据腾讯控股有限公司发布的综合业绩报告，截至2018年3月31日，其旗下移动社交网络应用“微信”的全球月活跃用户已达10.4亿^[2]。来自国际知名证券分析师和投资银行家玛丽·米克尔(Mary Meeker)的互联网趋势报告显示^[3]，中国用户每天花在移动应用上的时间达31亿小时，其中使用微信的时间占据了9亿小时。

在数字经济日益壮大的大环境里，移动社交网络是最大的受益应用之一^[4]。微信作为中国互联网最具代表性的产品之一，对数字经济发展起到了愈发重要的作用。微信定义的封闭好友关系概念，提高了社交网络中好友关系的强度；微信推出的在线文本通讯、红包、音视频通话、群聊、朋友圈等功能

能，聚合了数字经济用户；微信公众平台、企业微信、小程序等独特功能，帮助企业及用户进行营销及运营服务。企业可以凭借大数据、云计算、人工智能等技术，分析用户的购买兴趣，制定最有效的营销策略。

为了使经济效益最大化，营销平台通常采用计算的方式制定营销策略，这种方式被称为营销计算。营销计算需要综合数据挖掘、统计学习、机器学习等多方面的技术。然而，在移动社交网络广泛应用的大背景下，制定营销策略需要面对两项挑战：第一，如何利用海量同构网络的数据挖掘用户的行为模式；第二，如何融合多种异构网络的数据刻画更为精确的用户画像。

针对这两项挑战和实际场景中的各种问题，研究机构提出了许多行之有效的模型。例如，影响力最大化算法及其各种变体被用于辅助企业或个体经销商选择初始广告投放对象，以期望广告在移动社交网络中的传播影响得到最大化；通过特定的模型将用户映射为高维空间的数值特征向量，这些向量可以用于描述社交网络结构及构建用户画像，进而应用于潜在好友关系推测、多社交网络用户对应等任务；同理，学习内容在高维空间的数值特征向量在内容质量评估、个性化推荐等场景具有重要的意义。

结合最前沿的算法与模型，许多企业开展了基于社交网络的营销计算咨询服务。例如，深圳市兔展智能科技有限公司¹旗下的斐波那契数据便是一

¹ <http://www.fibodata.com>。

一个基于微信的社会化传播数据分析平台。该公司致力于营销内容的生产，每日可帮助用户推出逾万项营销内容，并可监测到日均数千万条传播流记录。

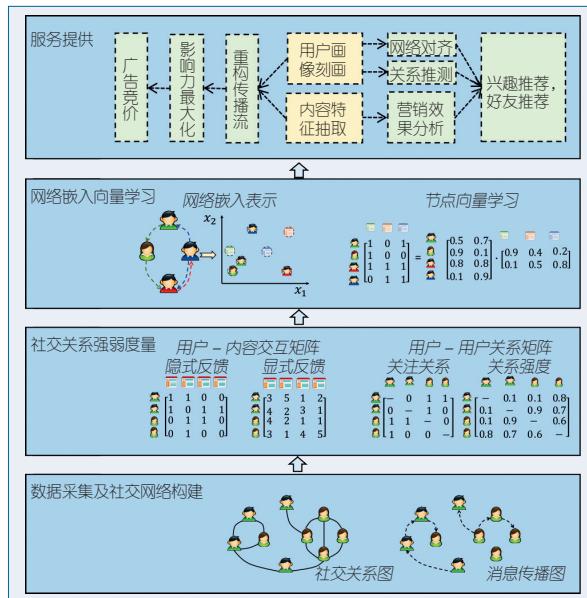


图 1 移动社交网络挖掘的基本框架

图 1 给出了移动社交网络大数据挖掘的基本框架，包括数据采集及社交网络构建、数据分析及社交关系强弱度量、网络嵌入向量学习、服务提供四个环节。在**数据采集及社交网络构建层面**，我们可以利用官方提供的接口，结合 URL 请求参数的形式获得用户的非敏感信息及传播流记录，重构一个与营销强相关的社交网络（即社交网络中的每一个节点都参与到了信息的传播中），特别是研究在社会大事件中（春运、国庆）传播与营销的关系。在**社交关系强弱度量层面**，我们可以通过用户的阅读、频率、评价等行为信息刻画网络节点间多维度关系的强弱。在**网络嵌入向量学习层面**，我们可以通过特定的模型从稀疏的异构图结构中学习到每个节点的数值向量表示，这些向量既可以表示节点的兴趣信息，又可以描述网络的结构特征。这些数值向量是在特定任务、特定目标下通过机器学习方法训练得到的，故在特定场景下的营销计算中具有卓越的效果。在**服务提供层面**，这些被学习的向量用于在线或离线匹配、识别、推荐等任务，以保证高速、

高效的服务质量。

社会大事件在移动社交网络营销中的体现

在营销计算中，对用户进行空间和时间上的分布分析是十分必要的，有助于根据社会大事件制定相应的营销策略，提高营销影响力。对于营销平台而言，获取社交网络中完整的信息是不现实的，故需要对能够获取到的局部数据进行统计分析，进而估计全局信息。

对数据的局部观测可以看作是在完整社交网络上的随机采样，这种采样是无偏的，即当采样获得的数据样本足够多时，可以认为样本中隐含的空间、地理、用户兴趣等多维度信息的分布与完整社交网络中对应维度的信息分布是相同的。本研究组与深圳市兔展智能科技有限公司开展合作，获取了 2016 年春运期间和 2017 年国庆、中秋假期期间几十万项 H5 轻应用（即 HTML5 语言生成的适配移动端的网页）所对应的传播纪录。该数据已经过脱敏，每一条记录描述了一个用户阅读一项 H5 轻应用的行为，包括阅读发生的时间、用户所在省份、该轻应用的转发人等。结合统计分析方法，我们可以有根据地凭借收集到的局部数据，估计全国范围内的人群分布。



图 2 2016 年春运期间实际到达人数前 10 省份到达人数比例及移动社交网络中这 10 省的阅读量提高比例

春运、社交与营销：春运是我国在农历春节前后发生的一种大规模人群迁徙活动。近年来，每年的春运都有数十亿人往来于全国各地，在人们享受古老、富有文化内涵而又极具生命力的节日的同时，新时代多样化、个性化的消费方式迎来了新的营销契机。通过对2016年春节前夕和春节假期末期的数据分析，我们发现，省际信息传播流量比重的差异可以反映出流动人口的地理分布。我们根据中国铁路总公司发布的《2016年中国春运大数据》绘制了春节前（截至除夕当天）到达省份人数最多的10个省份的到达人数比例（图2中蓝色柱形所示）。我们还观测到2016年除夕当天，这10个省的用户在被监测的微信朋友圈帖子上的总阅读量比例，相较于2016年1月前半个月的比例有所提升（如图2粉色柱形所示），与到达人数的比例十分相似。这说明了现实生活中的人口流动与线上传播流的变化具有同质性，故可以根据线上观测的传播流情况推断线下潜在目标群体的分布，进而制定合理的营销策略。例如，在2018年春运期间，福特公司²为旗下产品福睿斯汽车在上海往返成都的高铁线路上投放广告。广告主打亲情、放松、精准，福特公司提前覆盖春运这个时段，把握了高性价比的机会。

国庆、社交与营销：国庆是我国每年下半年的唯一法定长假，是许多人选择外出旅游的黄金时段。根据中国旅游研究院报告的数据，2017年国庆中秋假期共接待中国游客6.63亿人次³，带来了巨大的旅游、购物的商贸机会。图3展示了国家旅游局数据中心发布的2017年国庆中秋假日旅游消费热门目的地数据，颜色越深表示消费水平越高。我们统计了2017年10月1~6日来自各个省份的阅读量，并计算了相较于节后10月9~31日的平均值的变化量。图3中蓝色柱形表示阅读量提升，粉色柱形表示阅读量下降，柱形的长度表示阅读量变化的比例。我们发现，各省阅读量的变化比例与消费热度呈正



图3 2017年国庆中秋假期各省消费热度及移动社交网络中帖子观看量提升比例

相关关系。这说明，营销平台可以在节假日期间通过观测量的空间分布调整所投放的广告类型及选择的投放对象，达到更好的营销效果。例如，2017年国庆中秋假期期间，高德地图⁴推出了“吃喝玩乐大富翁”活动，基于用户的真实定位推荐周边优质兴趣点(Point-Of-Interests, POI)，让用户能够在不同的区域玩游戏，获得不同的福利奖品，也得到了良好的转化。数据显示，2017年“十一”期间，高德地图导航规划服务次数共近20亿次，驾车导航总里程超30亿公里，用户相比2016年“十一”同比增长超过100%。

构建异构网络：以深层知识定策略

微信围绕核心社交服务，建设朋友圈、公众号、移动支付、“小程序”等各方面的功能，形成了生态化的能力。微信将这些生态能力向个人或组织开放，使得其他主体获得技术和服务支撑，创新产品和服务，衍生出众多的小生态系统和新兴价值。为了在营销中更有效地利用这些移动社交网络的新功

² <https://www.ford.com.cn/>。

³ <http://finance.eastmoney.com/news/1355,20171010783602525.html>。

⁴ <https://www.amap.com/>。

能，营销平台需要挖掘蕴含在社交网络背后深层次的信息，以此获得无法通过简单统计方法得到的营销依据。

为了更好地描述社交网络的信息，我们希望将每个节点转化为一个数值向量，使得所有关系、结构都可以被量化描述。我们通过给定的任务目标和恰当的机器学习方法确定向量的具体数值，并根据定义在向量上的操作完成特定任务。假设移动社交网络 G 中有 N 个用户，涉及 M 个多媒体帖子。在维度为 d 的向量空间内，令 $d \times N$ 的矩阵 \mathbf{U} 表示用户的网络嵌入矩阵，其中第 i 列的列向量 \mathbf{u}_i 表示第 i 个用户的特征向量；令 $d \times M$ 的矩阵 \mathbf{V} 表示帖子的网络嵌入矩阵，其中第 j 列的列向量 \mathbf{v}_j 表示第 j 个帖子的特征向量。在给定的数据集 D 上，我们定义一个目标损失函数 $L(D; \mathbf{U}, \mathbf{V})$ ，希望学习所得的向量可以使目标函数描述的指标值达到最优，如平均误差最小化、最大概率似然等。我们定义函数 $f(\mathbf{u}_i, \mathbf{v}_j)$ 和 $g(\mathbf{u}_i, \mathbf{u}_{i'})$ ，分别用于判断第 i 个用户对第 j 个帖子的兴趣和第 i 个用户和第 i' 个用户之间潜在的联系。

重构传播流及影响力最大化



图 4 移动网络中的信息流传播应用

模拟传播流对于营销而言是一项非常重要的任务，用于对营销效果的提前估计。如图 4 所示，提前估计营销效果通常包含三步：学习节点表示、计算传播概率、影响力最大化决策。通过对历史帖子传播流的重构，我们可以学到社交网络中用户间针对特定信息发生传播的可能性。我们提出了一种通过异构网络学习特征向量重构

传播流的模型⁵。首先根据某一特定类型信息的传播流记录抽取转发网络、阅读网络和好友网络，构建异构网络。然后依据转发行为和阅读行为将用户的网络嵌入矩阵 \mathbf{U} 拆分成转发行为嵌入矩阵 \mathbf{U}^+ 和阅读行为嵌入矩阵 \mathbf{U}^- ，即每个用户对应两个向量 \mathbf{u}_i^+ 和 \mathbf{u}_i^- 。于是，可以定义目标函数⁵ $L(D; \mathbf{U}^+, \mathbf{U}^-, \mathbf{V}) = \sum_{(i, i', j) \in D} (f(\mathbf{u}_i^+, \mathbf{v}_j) + f(\mathbf{u}_{i'}^-, \mathbf{v}_j) + g(\mathbf{u}_i^+, \mathbf{u}_{i'}^-))$ ，其中函数 f 使用点积相似度， g 使用余弦相似度，我们希望此目标函数最大化。在模型参数确定后，我们可以使用 g 函数计算任意好友间传播当前类型信息的概率。

在移动社交网络平台中存在许多自媒体个人或团体，这些自媒体通常拥有规模庞大的粉丝群体及较高的影响力，可以参与到营销内容的传播任务中。为了使传播的范围更为广泛，我们希望调动更多个人用户参与到传播中，这需要我们利用模拟传播流的方法选择社交网络中的意见领袖，使传播覆盖人数的期望值最大化^[6]。给定一个选定的初始种子用户群 $S \subset \{1, \dots, N\}$ ，代表营销平台可以通过某种方式说服这些用户帮助转发传播信息。利用学习得到的概率传播图和独立级联模型，我们可以计算当前种子用户群的传播影响力 $\sigma(S; G)$ ，定义为可覆盖用户的期望数目。有理论证明，求解影响力最大化的种子用户群问题具有次模性质，可以使用差量贪心方法取得相较于最优解至少 63% 的数值。具体方法是，在一个选定的种子用户群的基础上继续加入种子用户时，选择使得 $\sigma(S \cup \{i\}; G) - \sigma(S; G)$ 数值最大所对应的用户。在这种方法的基础上，一些启发式算法根据传播树的特征进一步优化，使得选择的种子用户群在实际传播中更为有效。

用户向量学习

在移动社交网络应用飞速发展的大环境下，每个用户都处在多个移动社交网络中，并且在每个社交网络中表现出不同的特点或兴趣偏好。从网络中

⁵ 通常还需要加入负采样和正则化项使模型更为鲁棒，并使用平均误差最小化的方法将最大化问题转化为最小化问题，便于使用梯度下降法更新参数。这里为了便于理解，仅指出概念上的目标函数。下同。

抽取一个用户的特征向量^[7], 可以用于好友推荐、网络对齐、关系挖掘、社团发现等任务, 如图5所示。综合一个用户在多个网络中的特征, 是使得广告营销更有针对性的一种理想方式。然而, 许多用户不会轻易暴露自己拥有的不同应用的账户, 使得社交网络对齐成为一个严峻的挑战。一种有效的方式是利用用户有意或无意间暴露的信息进行对齐, 例如手机号、姓名、生日等信息。但在许多实际场景中, 这些元数据是无法完整获得的, 只能精确对应部分用户。在这种情况下, 使用用户特征向量的方法变得十分有效。

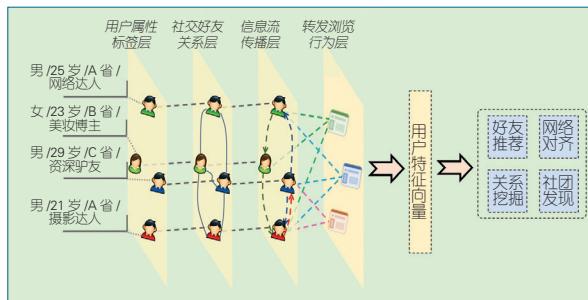


图5 移动社交网络中的用户嵌入向量学习

为了解决这个问题, 一种流行的方法是用户特征向量对齐。设两个需要对齐的社交网络图为 G_1 和 G_2 , 对应的用户网络嵌入矩阵分别为 \mathbf{U}_1 和 \mathbf{U}_2 。两个社交网络的社交关系数据集记为 D_1 和 D_2 。我们还拥有一个局部关系映射集 F , 记录了 $|F|$ 条 $\{i_1, i_2\}$, 表示 G_1 网络中第 i_1 个用户和 G_2 网络中第 i_2 个用户被确定是同一个用户。根据这个映射集, 我们希望匹配两个网络中其他潜在相同的节点。首先, 我们在两个社交网络内分别学习用户网络嵌入向量, 即使 $L(D_x; \mathbf{U}_x) = \sum_{(i,i) \in D_x} g(\mathbf{u}_{x,i}, \mathbf{u}_{x,i})$ 最大化, $x \in \{1, 2\}$ 。然后, 我们希望学习到一个映射函数 h 将两个社交网络中的向量对应起来, 并尽可能减少这个映射带来的误差, 即使 $L(F; \mathbf{U}_1, \mathbf{U}_2) = \sum_{(i_1, i_2) \in F} \|h(\mathbf{u}_{1,i_1}) - \mathbf{u}_{2,i_2}\|^2$ 最小化。与此同时, 我们试图构建一个解码器 $Q(\mathbf{u}_{2,i_2}) \rightarrow i_2$, 将 G_2 中学习得到的用户嵌入向量解码成为用户的编号索引。最后, 我们可以使用 $Q(h(\mathbf{u}_{1,i_1}))$ 来获得两个网络中的用户对应关系。模型中的几部分参数可以使用交替训练的方法, 即每次固定所有参数,

将一部分参数设置为可变进行学习, 使得模型的参数更加鲁棒并可以更快速地收敛。这里使用了迁移学习的思想, 实验证明这种方法的效果是非常理想的。

内容向量抽取

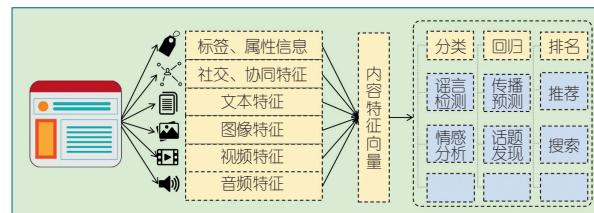


图6 移动社交网络中的内容嵌入向量学习

在移动社交网络中, 内容常以H5应用的形式呈现。H5应用是文字、图片、视频等多媒体元素的混合, 往往很难通过纯内容挖掘判断用户的关注点, 于是协同信息便是很好的依据。提取内容特征的方式及相关任务如图6所示。内容的网络嵌入向量 V 可以通过移动社交网络中的浏览行为构建出来。为了进一步提高准确率, 在学习 V 矩阵的时候需要引入一些指导信息, 以便向量有针对性地反映特定任务下的特征。例如, 我们提出了一种在移动社交网络中判断内容生存期的模型^[8]。在微信的生态中, 内容或网页通常由公众号创造或引入。最早接触新进入微信生态圈中内容的用户, 是关注公众号的用户。这些用户是否转发, 决定着是否能有更多人看到这条内容。利用这一特点, 我们在目标函数中引入相应的限制。我们建立一个刻画用户转发公众号内容意向的网络嵌入矩阵 A , 其中第 i 列表示第 i 个用户的转发意愿向量 a_i 。设从公众号转发第 j 条内容的用户集合为 B_j , 则这项任务对应的目标函数为 $L(D, B_1, \dots, B_M; \mathbf{U}, \mathbf{V}, \mathbf{A}) = \sum_{j=1}^M \sum_{i \in B_j} f(a_i, \mathbf{v}_j) + \sum_{(i,j) \in D} f(\mathbf{u}_i, \mathbf{v}_j)$ 。我们使用神经网络 $Q(\mathbf{v}_j) \rightarrow T_j$, 以内容的网络嵌入向量为输入, 输出内容的流行时长 T_j 。

挑战与机遇

社交网络竞价问题

在移动社交网络中投放广告已经成为越来越多企业的营销选择，尤其是雇用有影响力的人或自媒体进行营销，可以显著提高广告的展示量和宣传效果。然而，出于营销成本的考量和移动社交网络平台的限制，将广告投放给所有用户是不现实的。在挖掘意见领袖的基础上，营销平台还需要考虑经济层面的影响。在传统搜索引擎中，广告通常根据广告商次价竞拍的排名按序展示，有理论证明这种竞价方式是公平且稳定的。与此不同的是，在移动社交网络中，营销平台通常需要利用手中有限的资源为不同的广告内容进行营销推送，这些广告自然产生了竞争关系。另外，许多内容的客观价值难以简单估计，不合理的定价模式势必无法延续企业和营销平台之间的合作。一个正在兴起的领域是利用多代理强化学习方法解决这个问题^[10]。具体而言，移动社交网络可以被看作强化学习的环境，每个广告被看作一个独立进行决策的代理。每个代理需要根据自己期望的展示量、有效期、竞品的情况、违约金等因素进行后台“竞价”，最终营销平台可以根据后台的竞价排名进行展示，并进行相应的收费。这种方式的目标是在尽可能满足所有广告需求的前提下，使营销平台的收益达到最优。

社交推荐

推荐是营销的重要手段之一。结合挖掘到的用户兴趣画像和内容的特征向量，营销平台可以进行个性化的广告投放，从而提升广告投放的效果。通常，我们可以优化目标函数 $L(D; \mathbf{U}, \mathbf{V}) = \sum_{(i,j) \in D} f(\mathbf{u}_i, \mathbf{v}_j)$ ，使得模型既能描述用户的兴趣历史，又具有一定的泛化能力。于是，营销平台能够根据 $f(\mathbf{u}_i, \mathbf{v}_j)$ 选择分数最高且没有出现在历史中的若干内容推荐给用户。一直以来，如何进行更为准确有效的推荐是工业界、学术界的热点问题。已有大量方法使用协同或语义上下文等思想提高推荐的精确度，同时已有很多针对特定应用场景的模型。可以预见的是，未来将会有更多推荐算法服务于推荐任务。近年来越来越多的营销平台意识到，频繁推荐相似的内容可能会引起用户的“兴趣疲劳”，离线表现优秀的模型

在线效果会明显下降。一种合理的解决方法是，在利用用户历史兴趣进行推荐的同时，再探索一些新的、用户从未尝试过的内容^[10]。我们希望模型能够自适应地在“利用”与“探索”之间找到平衡，在给出的前 K 推荐中包含两方面内容，并设定合理的排序方案。这也是营销平台的一项重要任务。 ■



张远行

北京大学信息科学技术学院博士生。主要研究方向为计算机网络与表示学习。
longo@pku.edu.cn



边凯归

CCF 专业会员。北京大学信息科学技术学院副教授、网络与信息系统研究所副所长。主要研究方向为移动计算、计算机网络与大数据分析。
bkg@pku.edu.cn



宋令阳

北京大学“博雅”特聘教授，北京大学信息科学技术学院现代通信研究所所长，国家“杰出青年科学基金”主持人。主要研究方向为无线通信与机器学习。
lingyang.song@pku.edu.cn

其他作者：李晓明

参考文献

- [1] Bakshy E, Dean E, Yan R, et al. Social influence in social advertising: evidence from field experiments[C]// Proceedings of the 13th ACM Conference on Electronic Commerce. ACM, 2012:146-161.
- [2] 腾讯控股有限公司 . 腾讯公布 2018 年第一季度业绩 [R]. 腾讯 , 2018.
- [3] Mary M. INTERNET TRENDS 2018[R]. Kleiner Perkins, 2018.
- [4] Zhang Y, Li Z, Gao C, et al. Mobile Social Big Data: WeChat Moments Dataset, Network Applications, and Opportunities[J]. IEEE Network, 2018, 32(3):146-153.
- [5] Zhang Y, Bian K, Chen L, et al. Early Detection of Rumors in Heterogeneous Mobile Social Network[C]// IEEE Third International Conference on Data Science in Cyberspace. IEEE, 2018:294-301.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

盘活大数据资产 ——个人大数据权益保护与合理使用

段旭良¹ 郭兵¹ 吴帆² 等

¹ 四川大学

² 上海交通大学

关键词：个人大数据 数据权益 数据流通 资产管理

引言

数据是一种具有价值属性的宝贵资源，在信息时代更是成为个人、企业甚至国家的重要资产。大数据时代，许多面向个人消费者开展业务的公司、机构都在努力采集个人数据，将看起来是碎片的数据汇总起来，由此得到用户的联系方式、人际关系、家庭、习惯、消费、行踪等。“他们”如饥似渴地剖析我们，对我们进行画像，而我们却对自己的数据被谁用、怎么用毫不知情。大量个人数据的汇聚形成个人大数据，当前的普遍情况是对个人大数据的采集、处理和使用均笼罩在所谓的企业商业秘密之下，整个过程对个人和监管方都是不透明的^[1]。我们作为社会人，只要使用网络，接触社会，就没有绝对的隐私，而且面临着很多因个人信息泄露带来的安全威胁。

个人数据是一种宝贵的资源和资产，在数据时代是创新的源泉，其使用具有两面性。一方面，我们可以得到更好的个性化服务和体验；另一方面，也可能每天忍受着各种垃圾信息的轰炸，甚至遭遇真假难辨的电信诈骗。技术层面的信息安全和个人隐私保护技术已经研究了几十年，国内外个人信息保护相关法律法规从无到有，力度由弱变强，而泄露或非法获取个人信息的“黑色产业链”却愈演愈烈。一起起因诈骗酿成的悲剧令人扼腕叹息，但猖獗的

个人信息黑色产业链也从侧面证实了个人大数据潜藏的巨大价值，暴露了个人数据管理、使用和流通中存在的影响未来大数据产业发展的诸多问题。

管理和使用中存在的问题

个人数据管理和使用中存在的种种乱象是黑色产业链生存壮大的重要土壤，要想找出其中的问题，首先需要厘清个人大数据的概念。移动互联网时代，我们每个人每天都在直接或间接地生产和使用大量的数据，浩如烟海的数据中，哪些数据是属于我们个人的呢？哪些数据在“个人大数据”的范畴之内呢？

1998年英国颁布的《数据保护法案》中比较全面地界定了个人数据的内涵和外延，认为个人数据指“一个活着的自然人的数据集合，通过这些数据，或者这些数据和使用者占有的其他信息的组合可以辨识该人”。有关个人观点的表述及涉及到个人数据的使用时，使用者或其他人的意图也属于个人数据范畴^[2]。我国《网络安全法》中对“个人信息”的定义是“以电子或者其他方式记录的能够单独或者与其他信息结合识别自然人个人身份的各种信息”。欧盟的《通用数据保护条例》(General Data Protection Regulation, GDPR)对“个人数据”的定义是“任何已识别或可识别的自然人（数据主体）相关信息”。一般我们认为，个人大数据是个人生活

和工作活动中产生的、个人可以拥有或控制的数据，其中大部分是原始数据，其数据来源复杂、形式多样，将是我们每个人呈指数级增长的资产^[3]。

数据产权模糊、用户权益被侵害

个人大数据作为一种资源和资产，其所涉及的经济属性问题，虽然目前国内外都没有合理的解决方案，但随着数据经济的进一步发展，个人数据的所有权、隐私权、知情权、使用权、收益权、交易流通等问题是无法长期回避的。

产业界不是不注重个人数据，而是在有意回避，目前行业潜规则是“谁采集，谁拥有”，通过很多对用户来说流于形式的“告知许可”协议，各种互联网企业、平台、服务提供商、公共服务机构以及政府部门，以类似原始资本积累的圈地模式吸纳用户、积累资源，大大小小的运营平台对个人数据安全管理的层次也是参差不齐，出售和利用个人数据甚至个人隐私获利，“电信诈骗”等侵害用户权益现象时有发生，用户对个人数据缺乏甚至根本没有所有权、隐私权、知情权，维护用户权益极其困难。可以说，数据产权模糊是用户各项权益得不到保护的根本原因。

管理散乱、开放流通困难

大数据时代，我们每天的工作和生活都在有意或无意地产生大量数据，这些数据被各种软件，如浏览器、杀毒软件、社交工具等记录和采集，很多数据（如使用习惯、位置信息等）往往在我们不知情的情况下被收集。这些数据分散在各个网络运营平台，导致个人数据的存储碎片化、管理复杂化，用户对自己数据的掌控程度极其有限，形成个人信息孤岛，甚至一些业务平台成为“数据黑盒”，用户的隐私和各项权益难以保障。

对企业和政府部门而言，个人数据采集和管理的散乱现状严重影响了个人数据的流通、分享使用和依法监管。中国工程院院士邬贺铨曾指出，目前我国一些部门和机构拥有大量数据，但宁愿自己不用也不愿在有关部门和机构间分享，导致信息不完

整或重复投资^[4]。对于互联网新兴产业来说，由于缺乏有效的数据开放流通渠道，数据平台的割据现象致使线上/线下等多维度的个人数据汇聚非常困难，各个平台基于自有数据采用“盲人摸象”式的方法拓展新的业务和市场，个人征信、消费金融、产品精准营销、精准医疗等新的增值服务难以实现，严重阻碍了个人大数据经济价值和社会价值的发挥。

解决方法

对于个人数据，是严格保护还是开放使用？若没有合理合法的数据资源流通规则和健康的流通市场，可能的收益越大，“数据黑市”就会越猖獗。

很多国家和地区都通过立法保护个人信息安全，规范个人数据的采集、存储与使用。2017年6月1日，我国首部《网络安全法》正式施行，明确保护公民个人信息安全，防止公民个人信息被窃取、泄露和非法使用。2017年10月1日起施行的《民法总则》第111条规定，“自然人的个人信息受法律保护，任何组织和个人需要获取他人个人信息的，应当依法取得并确保信息安全，不得非法收集、使用、加工、传输他人个人信息，不得非法买卖、提供或者公开他人个人信息”^[5]。2018年5月25日，欧盟堪称“史上最严”的《通用数据保护条例》正式生效，这是一部严厉翔实的保护用户数据安全的法律，其核心之处在于明确用户对自己的个人数据有绝对的掌控权，堪称对目前愈演愈烈的个人信息安全问题的一剂猛药。但是，“是药三分毒”，对个人信息过于粗暴严苛的保护可能会直接导致部分企业退出当地市场，或者取消某些难以“合规”的重要业务，扼杀一些极具创新的新业态，影响数据产业的健康长远发展。

在大数据时代，面对具有重要潜在价值的个人大数据资源，在使用和流通这个问题上，不能因噎废食，而应该探索合理合法的数据资源流通规则和方式，要“疏堵结合”，在保护个人数据所有权、隐私权、知情权、使用权、收益权，以及明确数据采集者、使用者权责前提下，权衡隐私保护与数据价

值最大化的平衡点，构建阳光下运营的个人大数据流通市场，压缩黑色产业链生存空间，形成健康的个人大数据产业链和价值链。

数据确权及溯源

产权是市场经济的基础，数据产权是数据经济的基石，对个人数据没有控制权、知情权是导致隐私泄露、电信诈骗频发的重要原因之一。个人大数据产权对现有互联网商业模式具有颠覆性，是数据时代重大而又困难的问题。文学作品、专利、软件、IP核等人类高智力成果都已实现资产化——所有权清晰、货币可计量价值，能给所有者带来经济利益，并可以用于企业入股和清偿债务等。个人大数据价值密度相对较低、分布范围广、数量大，是一种个人的低智力成果，与高智力成果一样，是和个人相关的、由个人产生或创造的，应该具有明确的归属权。

数据确权是指确定数据的权利人，即谁拥有数据的所有权、隐私权、知情权、使用权和收益权并对个人隐私权负有保护责任等。数据确权的关键在于确定数据的所有人，包括数据的原始产生者和数据交易后的拥有者两方，而这两方的确定与数据溯源具有直接的关系。

数据溯源是追溯数据的演进过程，包括数据来源、管理、使用、交易、更新维护、失效退出等数据全生命周期的变迁过程，以及引起这些变化的因素。溯源技术的难点在于异构数据的处理，随着时间的推移和应用的需要，将产生各种各样异构的数据，这类异构数据如何实现溯源，是困扰业界的一个难点问题^[6,7]。

数据溯源能有效解决数据在使用过程中的监管问题，保证数据来源可查、使用可查和流向可查，可有效压缩黑色产业链生存空间。解决数据溯源问题，对数据权利人确定具有十分重要的意义。

数据安全与隐私保护

在个人大数据管理和使用中，个人数据安全和隐私保护往往是第一位的，直接关系到个人的安全和各项人身权利。个人隐私一般是指对个人敏感且

不愿公开的信息，通常有信息隐私、通信隐私、空间隐私以及身体隐私等几类，隐私程度、范围因人而异，没有绝对的标准，也没有绝对的隐私。对于隐私的定义主要有两类观点：一是基于价值，将隐私视为一种人权和商品；二是基于同源，认为隐私是个人的思想、认识、感知和状态。隐私会随着生活经验而改变，也依赖于特定场景，是动态的、多维的^[8,9]。

在法律层面，个人隐私的范围有一些明确的规定，如中国《征信业管理条例》规定，禁止征信机构采集个人的收入、存款、有价证券、商业保险、不动产信息、纳税数额信息、宗教信仰、基因、指纹、血型、疾病和病史信息；美国《公平信用报告法》认可的个人隐私信息包括支票、储蓄和证券账户的信息，以及驾驶记录、犯罪记录、医疗记录、保险单、收入、种族、信仰、政治倾向。但是，经个人同意的除外，依照法律、行政法规规定公开的不良信息除外。



因此，个人数据不全是个人隐私，但目前社会舆论和个人认知上存在将个人隐私范围扩大化的趋势，认为个人数据在任何情况下都不允许被商业化利用。实际上，个人大数据对于社会管理、商业应用和个人服务都有巨大的推动作用与经济价值。在出于善意和合法程序下，在合理的隐私保护范围内，个人数据应该得到有效利用，隐私的过度保护将降低数据效用。个人隐私保护的难点在于，在满足个人隐私安全保护的前提下，实现数据价值的最大化，或

者在隐私保护与数据价值之间寻找数据权利人可以接受的平衡点。但在不同行业和不同应用中如何确定这个动态变化的平衡点，目前是一个棘手的难题。

隐私保护的好坏直接关系到数据开放流通的产业前途，良好的数据生态环境必须以尊重保护个人隐私为前提。360公司创始人周鸿祎提出的大数据时代保护个人隐私的三原则对未来产业发展具有很好的指导性^[10]：(1) 存储在不同服务器上的个人数据应该是用户的资产，用户对个人隐私数据都具有所有权。(2) 为用户提供信息服务的公司，有责任和义务对用户数据进行安全存储和传输。(3) 使用用户信息要让用户有知情权和选择权，遵循“平等交换、授权使用”原则，泄露用户数据甚至牟利，不仅是不道德的行为，而且是非法行为。

个人数据的计量计价

个人数据计量计价是收益分配的基础，数据计量与计价的难点在于如何体现出数据包含的真正价值，形成一个良性的数据交易生态链。数据计量的基本单位可以是数据包、查询、视图、数据元组或者是数据调用次数。计价方式一般包括绝对计价和相对计价^[11]。绝对计价即数据所有者直接标价或者按行业内的规定定价，相对计价则是根据市场历史交易成功的类似数据价格定价。当前，数据定价模型相关研究主要有付费定价模型(pay-per-use pricing model)、预订定价模型(subscription pricing model)^[12]、免费定价策略、基于使用的定价策略等^[13]。但这些定价模式普遍存在如下缺点：定价模型允许套利，模型假设所有数据集是等价的，数据由客户进行缓存，以及数据提供者没有设定价格的指导性建议等^[14]。

数据流通机制

数据作为一种生产要素，只有合理配置和流通才能发挥其价值。数据开放共享与交易交换是数据流通的两种主要形式，其中数据共享指政府、企业以及个人间免费共享数据的产权、使用权；数据交易交换指将数据作为商品的交易行为和规则设计，涉及产权、隐私权、知情权、收益权和使用权等复

杂问题。

建立数据流通机制的目的是在数据需求方与数据提供方之间形成高效的信息交互渠道，从而避免信息孤岛，同时建立数据溯源、确权、定价等机制，让数据流通在阳光下运营，用健康的市场规则及运作最大限度地挤压数据黑市的生存空间。

数据收益分配

个人大数据作为个人的资产和财富，是一种被个人拥有的生产资料，个人应该享有数据流通、使用过程中产生的收益。数据收益应基于用户有效的数据，运营商通过数据分析和增值服务产生的数据收入，用户按比例享受收益分配权。

用户数据收益一般可以分为数据利息收益和数据分成收益两部分，其中数据利息收益的计算对象为用户的有效数据资产，即通过清洗和审核的个人数据，影响数据资产利息的主要因素为个人信息完整性和真实性(即基本信息完整度)，以及各类个人数据的商用价值率(即数据的资产化率)，个人数据利息可按日或者按月进行计息结算。数据分成收益是基于个人数据形成的数据产品和增值服务，服务商与用户按比例分成后用户得到的那部分。

个人大数据资产化管理

大数据是人们获得新知识、创造新价值的源泉，但是当前大数据格局类似于第一次工业革命早期阶段情况，商业上的成功很大程度上基于旧的价值创造模式。要真正实现大数据革命，不仅要有更好的算法、更快的速度，最重要的是要有创新的商业模式^[15]。个人数据银行模式参照现代银行成熟的商业运营模式，实现个人大数据资产的集中化管理和服务，数据使用情况透明可控，数据运营收益与用户分成，个人数据资产的管理者与经营者充分保证数据所有者的各项权益，是一种颠覆当前个人数据使用黑箱的创新性经营模式。

个人数据蕴含的巨大价值促使学术界和产业界重新审视其作为一种资源和资产的价值属性。2010

年，世界经济论坛 (World Economic Forum) 启动一项名为“重新思考个人数据 (Rethinking Personal Data)”的项目，发布《个人数据——一种新资产的崛起》报告，将个人数据作为“最新的经济资源”，列入“新的资产类别”^[16]。2014 年在北京举行的一场大数据产业推介会上，阿里巴巴集团创始人马云提出“人类正从 IT 时代走向 DT(data technology) 时代”的新观点，“今后拼的是数据能够给社会创造多少价值，用数据挣钱才是未来真正核心所在”^[17]。可见，在 DT 时代，数据能“挣钱”，是一种推动社会发展的重要生产要素，是与土地、劳动、资金等一样的重要社会资源。

个人数据银行

个人大数据在推动社会生活进步以及产业发展中发挥了重要作用，是每个数据所有者的宝贵资产和财富。对于个人来说，自己的数据怎么才能“挣钱”呢？成熟的商业银行模式或许有助于解决这个问题。银行的本质是资金融通的中介机构，即收集暂时不用的闲钱和散钱，并将其贷出，转换为生产性活动的资本，收益与用户分成，即存款利息。我们的数据和银行里的存款一样，虽然可能没在我们手里，但是其所有权是属于个人的，是我们的宝贵资产。对个人大数据实行资产化管理，汇聚个人数据，加强个人对数据的实际控制权，保护隐私权、收益权，形成良性的数据应用市场，可能是未来个人大数据产业发展的一个重要趋势。

个人数据银行是基于银行个人货币资产的管理与运营模式，以保护用户个人数据的所有权、隐私权、知情权和收益权为核心，建立个人大数据资产的管理与运营综合服务平台，包括数据确权、汇聚、管理、交易与增值服务等功能。

个人数据银行与传统商业银行发展历程比较见表 1。个人数据是个人财产的一部分，类似银行里的存款，虽然形式上有差异，但本质上是相同的，因此个人数据资产也可以采用成熟的商业银行模式进行管理和运营，既可以实现个人数据的集中有效管理，又可以促进数据增值和有序流通，带来一定收益，实现的目标包括：

1. 在个人数据产权和隐私保护的前提下，个人让渡数据的使用权，即产权“换”使用权，为个人数据有序的社会化流通提供重要基础。
2. 实现个人数据的产权化、资产化、商品化、集中化、服务化、专业化和收益化，使得个人数据量化有用、授权访问且可有序流通，降低个人数据的交易成本，使得数据资源配置更加优化。

结束语

大数据时代，个人数据不但是个人的数据资产，也是经济社会发展的重要资源。数据银行架构实现个人大数据资产化管理、运营和服务，以保护个人数据所有权、隐私权、知情权、使用权和收益权为核心，将个人数据的使用、流通置于阳光下运营，

表 1 传统商业银行与个人数据银行发展历程的比较

传统商业银行	个人数据银行
(1) 个人钱币自我保存：古代社会经济整体不发达，个人财富有限，个人钱币大多自我保存。	(1) 个人数据的自我保存：PC 时代或以前，将个人数据以纸张、磁带、硬盘等单一孤立的存储形式保存。
(2) 钱庄：又称银号、钱铺，明朝中期诞生，主要功能是汇兑和存放款。	(2) 个人数据存储与管理平台：以云盘或网盘为主要形式，实现个人数据存储管理，部分平台有交易功能。
(3) 现代银行：近代诞生，主要功能是货币资产的集中托管和专业化的综合运营，实现安全、便捷、收益和普惠的金融服务，是货币资产的管理者与经营者。	(3) 个人数据银行：大数据时代，主要功能是个人数据资产的集中托管和专业化的综合运营，实现安全、方便、收益和普惠的数据服务，是个人数据资产的管理者与经营者。
(4) 互联网银行：依靠互联网技术，在线为客户提供全方位无缝、快捷、安全和高效服务的金融机构。	(4) 互联网银行：互联网金融关键是数据与资金的整合，数据银行未来是互联网银行功能的外延性发展，二者将趋于融合。

是尝试解决个人数据产权模糊等问题，化解当前个人数据使用流通中的乱象，保护个人大数据权益的一种新业态和新模式；将有力推动个人大数据领域相关政策法规的完善，规范个人数据的增值服务和交易行为，斩断个人数据流通黑色产业链，盘活个人大数据这一宝贵资产。 ■

致谢：

本项目研究得到国家自然科学基金重点项目(61332001)、国家自然科学基金项目(61772352, 61472050)、四川省科技计划项目(2019ZDZX0045, 2019ZDZX0010, 2018ZDZX0010, 2017GZDZX0003, 2018JY0182)资助，在此表示感谢。



段旭良

CCF 学生会员。四川大学计算机学院在读博士，副教授。主要研究方向为个人大数据、数据清洗等。

5025968@qq.com



郭 兵

CCF 高级会员。四川大学计算机学院教授，博导。主要研究方向为绿色计算、个人大数据等。

495130092@qq.com



吴 帆

CCF 专业会员。上海交通大学计算机科学与工程系教授。主要研究方向为网络经济学、无线网络、移动计算、隐私安全。

fwu@cs.sjtu.edu.cn

其他作者：沈艳、申云成、董祥千、张洪

参考文献

- [1] Richards N M, King J H. Three Paradoxes of Big Data[J]. *Social Science Electronic Publishing*. 2013, 41(3): 41-46.
- [2] Margolis A. Five misconceptions about personal Data: why we need a people-centred approach to “big” data[C]// EPIC Proceeding. Wiley, 2013: 392.
- [3] 郭兵, 李强, 段旭良等. 个人数据银行——一种基于银行架构的个人大数据资产管理与增值服务的新模式 [J]. 计算机学报, 2017(1):126-143.
- [4] 新浪科技. 邬贺铨：需制定大数据国家战略 [EB/OL], <https://tech.sina.com.cn/t/2013-12-16/10179009292.shtml>.
- [5] 王春晖. GDPR 个人数据权与《网络安全法》个人信息权之比较 [J]. 中国信息安全, 2018, 103(7):40-43.
- [6] Galhardas H, Florescu D, Shasha D, et al. Declarative data cleaning: language, model and algorithms[C]// Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, USA, 2001:371-380.
- [7] Grigorios Karvounarakis, Todd J. Green. Semiring-Annotated Data: Queries and Provenance[J]. ACM Sigmod Record, 2012, 41(3):5-14
- [8] 刘雅辉, 张铁贏, 靳小龙, 等. 大数据时代的个人隐私保护 [J]. 计算机研究与发展, 2015, 52(1): 229-247
- [9] 孟小峰, 张啸剑. 大数据隐私管理 [J]. 计算机研究与发展, 2015, 52(2): 265-281.
- [10] 中国新闻网. 周鸿祎警示大数据时代六大挑战倡议重塑信息安全“三原则” [EB/OL]. <http://finance.chinanews.com/it/2014/09-24/6626241.shtml>. 2016-9-8.
- [11] Shen Y, Guo B, Shen Y, et al. A pricing model for Big Personal Data[J]. 清华大学学报：自然科学英文版, 2016, 21(5):482-490.
- [12] 黎春兰, 邓仲华, 张文萍. 云服务的定价策略分析 [J]. 图书与情报, 2013(01): 36-41.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>



封面设计说明

本期主题为大数据共享与交易。插图以电脑为核心，电脑屏幕上的数据和电脑前的柱状图以及数据流都是大数据 / 金融数据相关的元素，同时电脑屏幕上的曲线从屏幕中伸出，代表大数据的共享和传输。

设计：SEEEKLAB（设计总监 田力）

安全多方计算与数据流动

关键词：数据隐私 安全多方计算 工程化 数据流动

安 瑞 谢 翔 孙立林
矩阵元技术(深圳)有限公司

数据流通现状

当前的互联网时代，全民都在生产数据。上至70来岁的老人，下至五六岁的小朋友，都在网络世界为互联网企业的数据池做出自己的“贡献”。我们在网络上的任何轨迹都是我们思想的行为投射，毫无疑问，掌握了数据池中的数据，就可以刻画出对应个体的影子甚至全貌。

这些年来不断出现的数据泄露事件，让大家都在关注网络时代的隐私安全问题。2015年凯悦酒店集团旗下众多酒店的用户信息遭窃取；2017年美国信用评估机构Equifax约有1.43亿用户隐私信息遭泄露；2018年初Facebook约5000万用户信息在未得到用户授权的情况下，被第三方数据分析机构使用；Google在未经用户允许的情况下，将数据直接开放给第三方开发者。种种事件表明，数据如此重要，但并不安全。我们在享受企业所提供的便利服务的同时，并未考虑会产生如此多的安全问题。

“数据孤岛”这个词也在很多行业被不断提出。在医疗行业，各个机构都有自己的数据，但出于政策或是其他原因，即使是同行业甚至是同一家企业不同的机构间都未能实现数据的彼此开放。这就会造成研究某种疾病或者开发某种新药的机构无法获

得全量数据，使诊断与研发的精确性大打折扣。在人工智能(AI)行业，有句行话：没有数据就没有AI。AI的发展与强化离不开大量数据的支撑，要使AI模型更优、更强，就需要更多的数据供其进行机器学习。但在目前的机器学习领域，数据拥有方不愿直接给出其所拥有的数据，模型拥有方也不会给出自己的模型，造成两者之间存在不可逾越的鸿沟，导致无法创造出更强大的AI。

由于网络安全问题导致个人数据的隐私无法保证，各国政府都制定了相关政策与法律来保护数据的隐私安全。国内的《网络安全法》《侵权责任法》¹等定义，买卖个人数据视为犯罪，数据隐私为个人权利；欧盟的《通用数据保护条例》(GDPR)指出，企业要明确告知用户所收集的数据以及如何使用这些数据，并且个人有权对自己的数据做任何处理；美国加州签署了《2018加州消费者隐私法案》(California Consumer Privacy Act of 2018/Assembly Bill No.375)²，法案中规定，消费者对自己的数据有知情权、处理权，企业只有在得到用户授权的情况下才能对用户数据进行处理。

那么除了政策和法律来约束，是否有技术手段能够保证用户数据隐私安全？安全多方计算(Secure Multi-Party Computation, MPC)，作为现代密码学领

¹ 《中华人民共和国侵权责任法》由十一届全国人大常委会第十二次会议审议于2009年12月26日通过，自2010年7月1日起实施。

² 2018年6月28日，美国加利福尼亚州议会在没有反对票的情况下通过，被称为美国“最严厉、最全面的个人隐私保护法案”，定于2020年1月1日生效。

域的一个重要分支，可用于保护协议执行各方的原始数据隐私。

安全多方计算简介

安全多方计算由姚期智教授于1986年提出^[1,2]，之后由Goldreich、Micali以及Widgerson扩展。MPC作为密码学的一个重要研究方向，经过持续的研究已有丰富的理论成果。MPC指的是用户在无须进行数据归集的情况下，完成数据协同计算，同时保护数据所有方的原始数据隐私。参与各方在数据保留在各自本地的情况下，执行共同的既定计算逻辑(算法)，得到计算结果。数学形式化语言描述为^[3]，有n个计算参与方，分别持有私有数据 x_1, x_2, \dots, x_n ，共同计算既定函数 $f(x_1, \dots, x_n)$ ，得到正确的计算结果。计算完成后，参与各方除了自己的输入数据和输出结果外，无法获知任何额外信息。

MPC协议满足的基本性质是：

输入隐私性：协议执行过程中的中间数据不会泄露双方原始数据的相关信息；

健壮性：协议执行过程中，参与方不会输出不正确的结果。

这两点保证了数据流通过程中所需满足的基本要求。

MPC 原理解释

MPC自1986年提出以来，在学术界已出现众多的研究成果，所采用的底层协议或算法有GMW、BMR、Garbled Circuits、Oblivious Transfer、Secret Sharing、BGW等，工程实现有SPDZ^[4,5]、LEGO^[6]、Sharemind、Fairplay^[7]等不同的计算框架。从业界的工程实践来看，目前采用较多的是加密电路(Garbled Circuits, GC)和不经意传输(Oblivious Transfer, OT)两项密码学技术，基于GC与OT可构建出通用的两方计算架构。

在此架构体系下，安全计算的过程为^[8~10]：两个计算参与方P1和P2，想共同计算 $f(x_1, x_2)$ ，这里 x_1 为

P1所拥有的数据， x_2 为P2所拥有的数据， f 为计算逻辑。首先，P1将 f 转换为相应的布尔电路C，其中C的每个门都有一个真值表表示门的输入输出。然后，P1对真值表进行加密处理，得到加密电路 \tilde{C} 。同时，P1也对其输入进行加密，然后将加密后的输入与加密电路 \tilde{C} 一同发送给P2。此时P2就拥有了 \tilde{C} 和P1的加密输入(对应于P1输入的标签)，但未被告知P1的加密过程，因此P2就无法获知该如何使用自己的输入。这时，P2通过与P1之间执行1-out-of-2 Oblivious Transfer^[11]协议来获得加密输入(即对应于自己输入的标签)。之后，P2使用两方的加密输入对加密电路逐个门进行解密，获得电路计算结果。下面将详细介绍这5个步骤。

布尔电路生成

假设 $f(x_1, x_2)$ 是需要安全计算的函数，将此函数转换为布尔电路C，满足对任意 x_1, x_2 ，有 $f(x_1, x_2)=C(x_1, x_2)$ 。理论上任意函数均可表示成布尔电路。

加密电路生成

P1将函数转换为布尔电路C后，通过加密这些真值表将C转化为加密电路 \tilde{C} 。

为了说明这些真值表是如何加密的，我们用示意图来直观地解释。P1将电路的每个门表示成如图1所示的真值表。

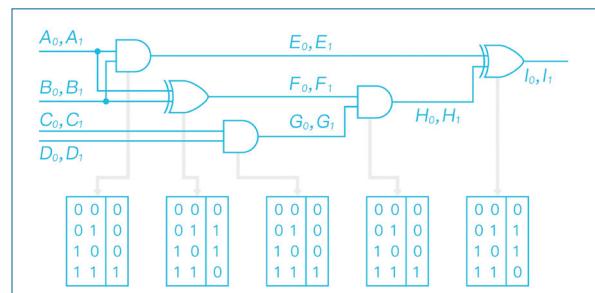


图1 将电路中的每个门表示成真值表

P1给电路中的每条线选取两个比特串长度为 k (安全参数，通常为128)的随机数标签 A_0 和 A_1 来代表0和1，如 A_0 表示0， A_1 表示1。对应关系只有P1知道，生成如图2所示的表。

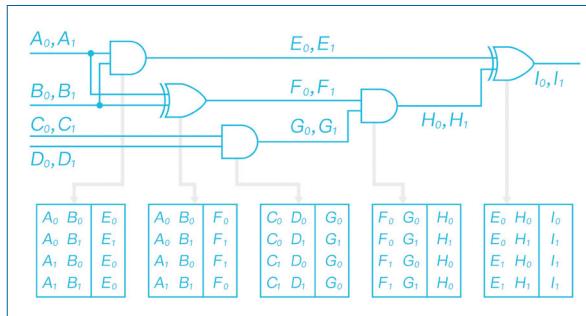


图2 选取随机数替代每个门中的0/1

P1 使用两个输入标签对每个门的输出标签进行加密，得到如图3所示的加密表，生成加密电路。可看出只有在获得输入标签的前提下，才能解密出加密表。

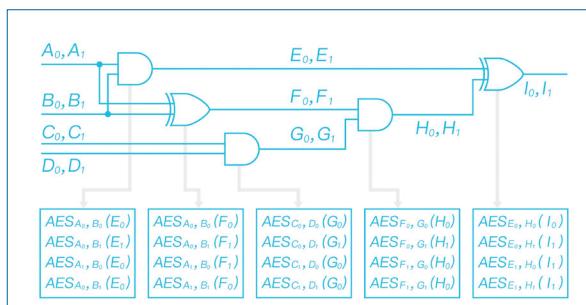


图3 使用输入线标签加密输出线标签

这里加密起两个作用：其一，每个加密都生成随机输出，使得输入与输出之间无法关联；其二，P2 在电路计算过程中不能获得除了应获信息外的任何其他信息，因为它不能获得对应输入的密钥。

P1发送与自己输入相对应的标签

生成加密电路后，P1 也需要对其输入 x_1 进行加密。首先 P1 将 x_1 转换为对应于电路 C 输入的布尔值，然后将布尔值的每一比特都用对应于 \tilde{C} 输入的标签 A_0, A_1 等来替换， x_1 的每一比特替换完成后，生成相应的标签。最后 P1 将这些标签以及加密电路 \tilde{C} 一同发送给 P2。

接收与 P2 输入相对应的标签

P2 此时已有加密电路 \tilde{C} 和 P1 的标签，还需要 P2 自己的标签来计算出计算电路。如前所述的加密

过程，P1 已经构造出 P2 输入的标签，却不知 P2 的实际输入。P2 知道自己的输入，但不能确定与输入值相对应的标签。这里采用 1-out-of-2 Oblivious Transfer 来实现。对于 P2 的每一个输入比特，通过与 P1 执行 Oblivious Transfer 来获得对应其输入的标签。这里，P1 是发送者，P2 是接收者，P1 输入为标签，P2 输入为 0 或 1，取决于实际的输入数据。这样 P2 就能知道与自己输入所对应的标签，同时 P2 既没有获得关于电路任何其他的信息，P1 也无法知道 P2 实际的输入数据。

Oblivious Transfer 解释：它是指发送者从一个值集合中向接收者发送单个值的问题，这里发送者不知道发送的是哪一个值，而且接收者也无法获知除了接收值之外的其他任何值。形式化描述为：发送者有由 N 个值组成的集合，接收者有索引 i , $0 \leq i < N$ 。协议执行完成，接收者只知道 N_i ，不知道 N_j ，这里 $j \neq i$ ，并且发送者不知道 i ，这称为 1-out-of-N Oblivious Transfer。

对于 $N=2$ ，即为 1-out-of-2 Oblivious Transfer，接收者在计算加密电路之前，首先根据自己的输入数据获得与之相关的标签，由于标签是发送者定义的，因此接收者与发送者之间执行此 OT 协议。接收者按照其输入的每个比特，以 $b=0/1$ 为输入，发送者以标签 X_t^0, X_t^1 为输入，协议执行完成，接收者获得标签 X_b 。协议执行过程满足以下性质：

1. 发送者不可获知接收者选择的是哪个标签；
2. 接收者无法获知另一个标签 X_{1-b} 。

释例：假设佩吉 (Peggy) 输入为 $a=a_4 a_3 a_2 a_1 a_0=01101$ ，她将 $X_0^{a_4}, X_1^{a_3}, X_1^{a_2}, X_0^{a_1}$ 以及 $X_1^{a_0}$ 发送给维克托 (Victor)，维克托无法知道佩吉输入的信息，因为标签是由佩吉随机生成的，对维克托来说就是随机串。为了计算出加密电路，维克托还需要对应于自己输入的标签。他针对自己输入的每一比特执行 OT 来接收标签。假设维克托输入为 $b=b_4 b_3 b_2 b_1 b_0=10100$ ，他首先针对第一个比特 $b_0=0$ 接收佩吉的两个标签 $X_0^{b_0}$ 和 $X_1^{b_0}$ 之间的一个。通过执行 OT，他接收到 $X_0^{b_0}$ 以及剩余比特所对应的标签。OT 执行完成后，佩吉不会知道任何关于维克托输入的信息，维

克托也不会知道其他标签信息。

电路计算

P2 收到自己的加密输入后，P2 逐个门依次进行解密，并将每个门解密后的输出值传递给下一个门，作为下一个门解密的输入。整个电路执行结束，P2 获得输出标签，如图 4 所示。而标签的实际对应关系 P2 知道，因此 P2 可获得最终计算结果，并将结果发送给 P1，这样双方获得计算输出结果。

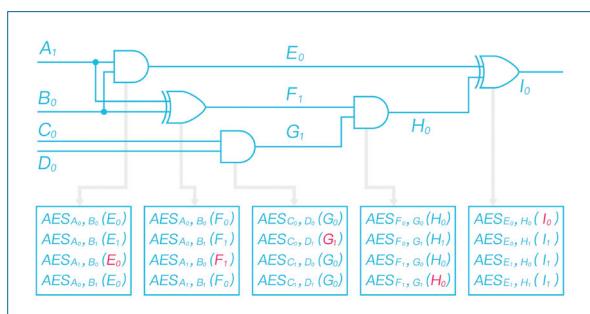


图 4 按门依次解密获得输出标签

工程化实现

MPC 发展至今已有相当成熟的理论研究，但是工程化实现道路仍任重道远。行业内有一些团队在做 MPC 的工程化实现，但是能够进行工程实现并研发出相关产品的企业在国内乃至全球都屈指可数。如何建立一个 MPC 产品化的范式，是 MPC 落地必须考虑的事情。

首先，针对上层不同的应用算法，如何方便开发人员直接在集成环境中开发？其次，对于不懂硬件的开发人员，如何使他们快速地将算法转换为电路？再者，进行应用开发时，如何让前端调用到底层的 MPC 算法？这些都是 MPC 实际落地与工程化时所要解决的问题。

为了降低开发者生成电路文件的门槛，通过 MPC 编译器可直接将应用算法代码转换成电路文件；在应用开发工程中集成 MPC-SDK，可连接到计算网络作为一个计算节点，并且通过 SDK 提供的 API 可为应用层调用底层 MPC 算法提供方便。目前，矩阵元技术有限公司正在做 MPC 虚拟机 (MPC-

VM)，如 C++、Java、Python、Go 等语言所开发的应用算法均可在 MPC-VM 中执行，极大地增加了 MPC 开发的可扩展性与易用性。这一整套开发工具与模块，使得行业的数据隐私计算场景正逐步落地。

MPC 应用场景

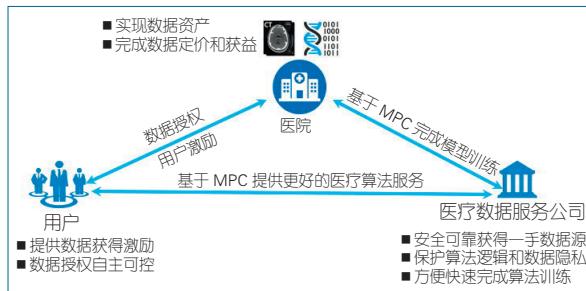
医疗行业

医疗行业如今有三大难题：立法漏洞、缺乏信任、隐私控制，均可归结为数据的隐私安全无法得到保证。

如今的医疗环境下，病患在医院的数据并不完全被患者持有，相反地，医院掌握着对病患数据的处置权。从医疗行业了解到，医院积压着大量的病患数据，包括诊断单、体检单、CT 影像等，个人从这些数据中除了了解到自己的健康状况外，也不会有其他用处。但医疗研究机构、医疗企业、制药企业对这些数据确是求之不得，因此用户个人的数据未经允许被滥用的情况时有发生。另外，医院本身诊断了各种病情，积累了珍贵的病情案例数据，但这些一手数据对于医院来说具有极大的重要性，在没有得到隐私保证的前提下，无法与其他企业和机构进行共享。现在互联网医疗、AI 医疗的兴起，大量的企业需要这些数据来帮助他们进行产品的研究，这些数据通过算法训练，来帮助实现更加快速的临床诊断，对精准医疗可起到更好的促进作用。

病患在医院进行肺病诊断时，医院通常都会对病患拍摄 CT 影像进行观察，因此医院积累了大量的 CT 影像数据。为了提高诊断的精确性以及效率，目前有很多医疗服务公司开发算法模型来帮助医院对 CT 影像进行识别分类，从而确定肺部的种类以及严重程度。但是由于医院数据的隐私限制，医院不能将数据透露出来。同时，医疗数据服务公司的算法模型是其自己研发的智力财产，也不会直接提供给医院使用。因此，MPC 在此场景正可得到应用，图 5 为基于 MPC 实现医疗数据分析的框架。

在此框架下，用户在医院所拍摄的肺部 CT 影

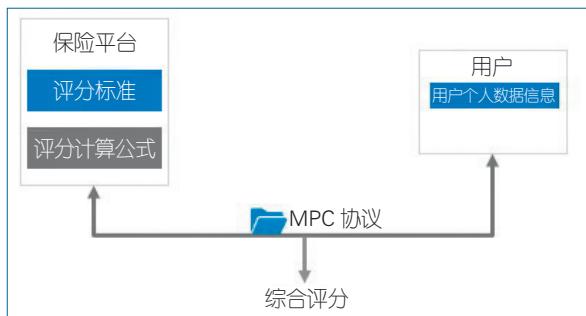


像数据归自己所有，医院只有得到用户的授权才能获得 CT 影像的使用权。通过授权机制，用户也可获得相应的经济激励。同时，授权的机制保证了个人数据完全自主可控。医院提供 CT 影像数据，医疗服务公司提供算法，医院与医疗服务公司各自在本地部署 MPC 节点，执行 MPC 协议，实现在保护各自数据以及算法隐私的前提下，完成对医疗企业算法模型的训练。医院根据医疗企业的数据使用需求来对数据进行定价，并获得相应的收益。由于医院可不断地提供最新的数据，那么企业的算法模型也能持续地改进，最终用户能够获得更精准更加高效的医疗服务。

保险行业

在保险行业的应用场景下，用户在保险公司进行投保时，保险公司会根据自己的用户模型对用户进行综合评分，根据评分对保单进行评估。图 6 为 MPC 在保险行业的应用框架。

以某保险公司为例，用户要对自己的健康进行投保，保险公司会根据自己的评分算法模型（包括评分标准和评分公式）对投保人进行评分，在进行



精准评分时，会收集用户各个维度的数据，包括年龄、收入、家庭、工作、过往健康状况、家庭病史、保险经历、性别、性格、所在地区等。保险公司先根据评分标准对用户某个单独项进行打分，如年龄 30，打 1 分；收入 50k，打 1 分；性别女，打 2 分；性格活泼，打 2 分。然后根据评分计算公式，如总分 = (年龄得分 + 收入得分) × 性别得分 × 性格得分，计算出最后的总得分。由于评分标准和评分公式为保险公司自己研发的算法模型，因此不会向用户透露具体是如何进行评分的。同时，用户关于自己的数据是其最为隐私的数据，也不愿直接告诉保险公司。通过在用户终端和保险公司各自部署 MPC 节点，在保证用户数据隐私以及保险公司评估策略隐私的前提下，共同执行 MPC 协议，计算出对用户的综合评分。根据所计算出的评分结果，保险公司可做出最好的决策，从而降低公司的风险，亦能够制定出适合投保人的最佳投保策略。

征信行业

用户通常在不同的征信机构中有不同的信用数据，目前最典型的就是人行征信与互金征信数据不通，用户在各家机构的信用数据无法进行整合。信用服务提供商获取来自不同数据源的信用数据，来建立信用评估模型具备相当大的难度。要想对用户在各个行业的信用进行全方位的征信评级，目前近乎难以实现。

目前的第三方信用评级机构对个人进行信用打分时，通过对用户在各个行业消费的数据进行整合，然后根据其信用评估模型计算出用户的信用分。毫无疑问，信用评级机构是以用户让渡隐私的前提，来对用户进行信用评级。为了解决信用评级中的隐私性问题，可采用基于 MPC 技术的解决方案，图 7 为基于 MPC 的分布式信用评估框架。

用户由于个人的行为，会在银行、不同商家、租房机构等不同的企业留下用户相应的行为数据。为了保证用户对自己数据的拥有所有权，企业只有在得到用户授权时才能使用其数据，可通过以区块链设计的分布式架构，使得用户与企业之间自主建

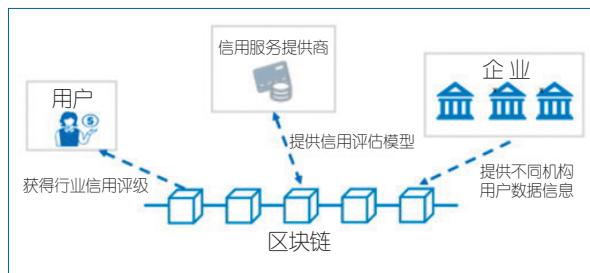


图 7 基于 MPC 的信用评估

立连接。信用服务提供商需要用户分布在各家企业 的信用数据，才能使用自己的信用评估模型对用户 进行评级、计算出信用分。显然，信用服务提供商 的评分模型不会透露给数据公司。这里，通过在信 用服务提供商与企业之间部署 MPC 节点，实现在 保护用户数据隐私的前提下，对用户进行信用评级， 并将 MPC 计算后的评级信用分上链存储，所有人均能 查询验证。

用户通过数据授权，获得数据使用收益；企业在 得到用户数据使用授权后，与信用服务提供商之 间通过 MPC 进行信用评级，保证了用户数据的隐私， 同时可利用评级结果为用户提供针对性的服务，改 善服务质量。信用服务提供商通过对用户多维度数 据信息的可信评级，获得名声经济以及企业的信赖， 并且在分布式架构下，其他信用服务提供商也可参 与竞争，使评级服务也能形成一个大的市场，促进 更加精准的评级。

联邦学习

人工智能为当前业界最热门的技术，毋庸置疑， 它已在很多行业有着广泛的应用。但是目前机器学 习面临着很大的瓶颈——数据获取困难。

机器学习算法需要消耗以及处理大量的数据来 学习关于人、业务流程以及各种事件的复杂模式。 当前，机器学习主要用于解决分类和聚类的问题。 算法在开发期间从训练数据中学习，在实际部署后， 也能从实际数据中不断学习，从而持续地改进算法 模型。机器学习算法的发展依赖于大量的数据供给， 为了拓宽和丰富算法所产生的相关性，机器学习需 要来自不同来源、不同格式以及不同业务流程的数

据。并且算法一旦投产，更是需要持续地学习大量 以及多样化的数据集，使得算法模型保持最新状态 并且持续成长。为了解决机器学习领域数据获取不 足的问题，已出现许多学术成果^[12,13]，在保证输入 数据隐私的前提下进行机器学习。

当前由于各家企业对自己数据的保护以及隐私 法案的限制，造成各家企业之间数据不通，形成了 一个个数据孤岛。联邦学习就是为应对数据隐私而 提出来的，它是在保证各数据源提供者数据隐私的 前提下，对机器学习模型进行训练增长的技术概念， 其核心为保证数据不出本地，同时实现模型增长。

在电商行业的持续扩张背景下，商户要想提高 市场影响力并争取到更多的客户，需要大量的用户 信息，并针对性地为其提供商品推荐。但用户的信 息一般分布在不同的企业，如社交平台拥有用户个 人偏好的特征数据，电商平台具有商品特点的数据 特征。但要保护企业数据安全，并且不能泄露用户 数据隐私，这造成了社交平台与电商平台之间的数 据壁垒；同时，不同平台的用户数据和用户特征数 据是异构的，联邦学习正可在此发挥关键作用。图 8 是基于 MPC 实现的联邦学习来解决此问题的框架。



图 8 基于 MPC 的联邦学习

由于电商平台通常均有用户的商品浏览记录， 因此会根据用户的浏览记录通过推荐模型算法，向 用户推荐一些商品，但获取到的用户数据维度不足， 无法为用户更加精准地推送商品。通过在两个不同 平台上部署 MPC 节点，构建基于 MPC 的联邦学习 架构。满足社交平台在保证用户数据隐私的前提下， 将数据分享给电商平台进行使用，并获得数据使用 收益。电商平台因为有更多的用户数据来源，使推 荐模型更加精准，从而增加商品的销售量。用户为 此可减少花费在电商平台寻找合适商品的时间，而

且买到更加适合自己的商品。

结语

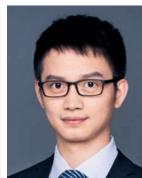
回首过往，互联网已走过 30 载，其发展有目共睹，任何行业恐无出其右。人工智能概念提出至今亦有 60 年，近几年开始蓬勃发展，但是目前人工智能行业面临着诸多问题。安全多方计算理论的建立至今不到 40 年，真正将 MPC 带向实际应用的是在 2015 年人类基因组的基因分析大赛上，MPC 被应用于基因安全分析，之后众多行业将 MPC 应用于不同的有隐私保护需求的场景。区块链的理论出现至今已有 10 年，比特币作为区块链代表性的应用，在全球各地的计算机上也运行了近 10 年。近几年，全行业都在寻求如何使用区块链来解决各自领域存在的痛点。

互联网发展至今几乎已形成几大互联网巨头垄断的形势，数据正被他们以一切可能的方式获取。在给人们生活带来便利的同时，能否做到保护用户的利益，成为当今互联网亟须解决的问题之一。人工智能能够让机器更好地为人类服务，MPC 致力于保护数据隐私，区块链是为数据带来价值的解决方案。我们相信，未来此三大技术并驾齐驱，才能将互联网真正带进下一个时代，实现数据的流动。 ■



安 瑞

矩阵元技术（深圳）有限公司安全算法架构师。
anrui@juzix.net



谢 翔

矩阵元技术（深圳）有限公司安全算法总监。主要研究方向为公钥密码学、格密码学和全同态密码学。
xiexiang@juzix.net



孙立林

矩阵元技术（深圳）有限公司 CEO。主要研究方向为密码学及零知识证明、安全多方计算、全同态加密、算法博弈论等计算复杂性领域。
sunlin@juzix.net

参考文献

- [1] Yao A C. Protocols for secure computations[C]//*The 23rd Annual Symposium on Foundations of Computer Science*. IEEE, 1982.
- [2] Yao A C. How to generate and exchange secrets[C]//*The 27th Annual Symposium on Foundations of Computer Science*. IEEE, 1986.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

2018 NOI 教师培训圆满收官 逾千名中学教师参加

教师是中学计算机科学教育的关键。2018 年，由 CCF 主办的 13 个教师培训分别在南京、北京、绍兴、长沙、大连、杭州、中山、成都、绵阳等 CCF 计算机科学教育基地顺利举行，共有来自全国 30 个省（市）713 所学校（单位）的 1079 名教师参加。

CCF 邀请了 30 多位 NOI 中学资深教练，从教师的实际教学问题和需求出发，针对数据结构、动态规划、组织教学和开展奥赛等知识难点及教学方法做了逾百场专题讲座，让教师通过近距离沟通和交流，解惑释疑、开阔思路，提升专业教学水平。

为了使经济困难的学校和教师有机会参与到提升自我的交流平台中，CCF 在每期培训中都会减免他们的注册费。CCF 将继续大力发展中学计算机师资培养，推广计算机编程知识普及，为中学计算机教师提供专业的交流和提升平台。

数据共享和交易 ——数据的质量、价值和价格

关键词：数据共享 数据交易 质量评估 价值评估

张 兰 李向阳 李安然 等
中国科学技术大学

数据共享和交易现状

随着信息化进程的加速，当今数据的体量正呈指级爆炸式增长。体量庞大、形式多样的数据已经成为个人和机构的重要资产，甚至是国家的战略资源，其丰富的内涵若得以充分挖掘利用将发挥巨大价值。在此目标的驱动下，人工智能和大数据科学技术正飞速发展。但与此同时，我们不能忽略大数据的一个稀有属性——协同作用，即多个数据集作为一个整体的价值要大于各个数据集价值的简单相加。当前，由于数据共享技术的限制、数据所有者的自私性及其对隐私安全的担忧形成了大量数据孤岛，这造成了数据资源极大的浪费，也严重阻碍了大数据充分发挥价值。

通常数据交易市场中涉及到买家（数据消费者）、卖家（数据所有者）和平台（数据代理商）三方，他们从各自的利益出发必然会问“数据质量如何？”“数据值多少钱？”和“数据卖多少钱？”这三个问题。实际上，在整个交易流程中准确可信的数据质量评估、价值评估和公平的数据定价是保障买卖双方权益、维护平台声誉、构建公平可信的规范市场、维护健康的数据共享和交易生态的关键问题，也是亟待解决的难题^[1,2]。

数据的质量和价值评估

当我们把数据作为一种有价资源进行买卖时，就需要对数据的质量以及价值给出一个合理的估计和准确的量化。质量评估关注数据内容本身的特性，而价值评估则是在评估数据质量的同时，进一步综合考虑数据在生产过程中的开销和在不同应用中的产出。尤其是数据的易复制、流通渠道难管控的特点使得售出数据难以实现退货，这进一步提高了在售前对数据进行准确可信的质量和价值评估的要求。买家在出钱购买前需要了解数据本身的质量及价值，以选择适当的数据和指导出价；卖家需要掌握数据的质量和估值，以指导售价，改善数据质量，提升数据价值；平台需要对数据质量和价值进行监管，以筛除质量低劣的数据，防止恶意报价扰乱市场，同时为用户推荐高性价比数据，从而提高用户的满意度和平台声誉，构建健康的市场生态环境。

面临的挑战

目前的研究工作普遍认为数据质量是一个多维度的概念，不同的工作提出了多种数据质量维度的定义和评估方法^[17]。综合现有工作，其中五个重要的维度得到了普遍认可，包括内在质量、表述质量、上下文质量、可访问性和可信赖性。其中，内在质量是基础需求，旨在衡量内容本身在容量、准确性、完整性、及时性、唯一性、一致性、安全性、源可靠性等方面的质量。该质量维度主要取决于数据源和数据收集、

处理的过程。表述质量是高级需求，与数据格式（简明和一致的表示）和数据含义（可解释性和易于理解）相关。上下文质量即数据必须与当前决策过程（任务）的上下文相关，适用于目标应用场景。由于任务及其上下文随时间和数据消费者的变化而变化，准确评估上下文质量和获得高上下文质量的数据是一项非常有挑战性的工作^[18]。可访问性即数据在多大程度上可供买方使用或检索，如通信成本和数据交付的延迟。可信赖性是指卖方及数据采集源的信誉和可信度。文献[17]指出，高质量的数据应该是内在质量良好，数据表示清晰，数据消费者可轻松访问，并且适用于具体的应用场景（任务）。

数据价值不仅依赖于数据质量的好坏，还受到成本、市场等多方面影响。传统无形资产的价值评估方法大体分为三类——基于成本、基于市场和基于收益^[3]。如采用**基于成本的思想**，数据的价值取决于收集、处理、存储的成本，但数据通常是作为信息系统的副产品生成的，生产成本与其他产品共享，所以难以确定数据的生产成本，且其生产成本也很难体现数据的真实价值；如采用**基于市场的思想**，数据的价值取决于同一市场上可比数据集的市场价格，但“可比性”的定义不明确，而且通常也不存在很相似的数据集，故难以准确估值；如采用**基于收益的思想**，数据的价值取决于买家能从数据中获得的总收益，该方法是主观的，仅在评估特定应用时可用，不同买方的估值可能会有很大差异。综上所述，目前尚缺少能准确评估数据全部价值的模型和方法。

面对数据交易市场上体量庞大、类型多样的数据，数据的质量和价值评估工作还面临以下挑战。

1. 多维质量量化评估：现有工作给出了数据质量丰富的评估维度，但很多维度仍采用定性分析，缺乏具体的量化模型和方法。尤其面临大量的非结构化数据，对数据内容进行分析是十分具有挑战性的问题，量化其内容的质量更是难上加难。

2. 数据集合质量评估：现有工作评估数据质量时大多针对单个数据单元（如一个文本、一张图片），缺乏对数据集合整体质量评估的方法，而实际

在数据共享和交易平台上分享或贩售的多为数据集（如1万个文本、10万张图片）。若简单通过单个数据单元的质量统计得到数据集整体质量，则忽略了数据单元之间的关系对数据集质量造成的影响。

3. 数据价值动态评估：不同模态和内容的数据，稀有程度和获取难易程度不同，如何合理评估这些因素，并结合数据质量形成数据整体价值的量化估计仍是一个难题。现有的数据评估工作通常关注于数据的静态质量，忽略了数据价值的动态特性，即随着数据采集存储设备的更新换代，数据挖掘模型方法的优化更新，以及应用场景和数据消费者需求的变化，数据的价值也在随之改变。这些动态因素进一步增加了数据价值估计的难度。

从五个维度实现综合评估

面对以上挑战，以数据集的合理质量和价值评估为目标，我们提出从以下五个维度实现从数据单元质量到数据集合质量，从数据内在质量到数据上下文质量，从静态质量到动态价值的综合评估。

1. 数据单元的内在质量。该维度衡量单个数据单元本身质量的好坏，是数据质量评估的基础需求。我们主要从数据的完整性、准确性和精密度三个方面来衡量数据单元的内在质量。完整性是指数据内容无缺失的程度，如表格的缺项程度。准确性是指数据正确、可靠的程度，如文本的拼写和语法正确的程度^[4]，图片及其标签的一致性等。精密度指数据的采集和存储的精度，如传感器数据的测量精度，图片和语音的清晰程度^[5,6]等。针对不同的数据类型，完整性、准确性和精密度的评估方法不尽相同。

2. 数据集的内在质量。该维度衡量整个数据集本身质量的好坏，除了计算集合中数据单元的完整性、准确性和精密度的统计值（如计算最小值、平均值、方差等），还需要考虑数据单元之间的相互关联，如数据的一致性和客观性。数据集的一致性是衡量数据单元之间逻辑上是否兼容。客观性是指集合的数据内容组成无偏见和公平的程度，有偏见的数据集可能传达误导或不真实的信息，例如采

样的数据集其内容分布应该与数据的真实分布相符。文献[1,7]指出，还应该考虑数据集的声誉，其反应了数据来源受信任或重视的程度，例如知名机构提供的数据比某不知名的机构提供的数据更值得信任。

3. 数据集的上下文质量。该维度衡量数据集在不同上下文中的价值，即使用价值。数据本身没有价值，只有在人们使用它时才会变得有价值。首先，数据集的高内在质量是其发挥大价值的基础。但由于数据协同作用，数据集的价值不是数据单元价值的简单加和，一个数据集的价值还取决于它的用途。对此，我们提出从五个维度衡量数据集的上下文质量，分别是**相关性、多样性、时效性、完整性和数据量的适度性**。**相关性**表示数据与任务相关或适用的程度，其衡量基于三个观察：首先在单个任务上，数据集越适用其使用价值越大；其次，数据集适用的场景越多其使用价值越大；最后，数据集适用的任务重要性越高其使用价值越大。**多样性**衡量了数据集包含信息量的多少，很多场景下我们都希望数据具有适度的多样性，例如多样化的数据集可以帮助机器学习模型减轻轻过度拟合并改善模型泛化能力。当然，不同任务对多样化程度的需求不同。例如，在人脸识别模型的训练任务中，我们希望数据集包含多人的不同角度的脸，而在屏幕人脸识别模型的训练任务中，我们则希望得到一个用户不同角度的脸。通常数据集**多样性**问题可以将数据单元抽象为高维空间中的点，用数据点间的最小距离、数据集的熵或空间覆盖问题来建模。与大多数其他资产一样，数据价值会随着时间的推移而贬值或升值，因此我们还需要度量数据的**时效性**来表达其在时间上适合任务的程度。数据价值改变的速度取决于数据的类型和内容。**完整性**^[8]表示数据的广度、深度和可扩展性方面满足当前任务的程度。它被定义为“给定数据集包含的能真实描述任务关心的目标物体数据的多少”。**数据量的适度性**表示数据的数量足以满足一项特定任务的程度。在很多实际应用中并非数据越多越好，适当的数据量取决于问题的需求、算法的复杂性、数据收集的成本及

计算资源等限制。

4. 数据集的流行度。该维度进一步反映了数据在动态市场需求下的使用价值。每个数据集都有其适用的任务，然而应用市场对不同任务的需求日益更新，从而使得数据集在不同市场需求背景下的价值也不断变化。这种变化可以用数据集的流行度来反映。例如随着基于图像深度学习应用的飞速发展，各类图像数据集得到了广泛的使用。可类比论文的引用量是其质量评价的指标之一，我们可以从多方面考虑数据集的流行度，如数据集被使用/引用的次数、被搜索的热度、被相关论坛讨论的热度等。

4. 数据集间的关系。该维度反映了多个数据集间的关系对给定数据集价值的影响。我们可以从稀缺性和协同性两方面来衡量数据集间的关系。稀缺性是指当前数据集与市场上其他数据集的相似程度，相似的数据集越少，数据集越稀有，相应的价值越高。协同性衡量数据集间的协作能力。如果给定数据集通过和其他数据集联合发挥的价值高于数据集个体价值之和，则该数据集具有较好的协同性。协同性越高的数据其潜在价值越大，被购买的可能性也越高。对数据集协同性的评估有利于用户和平台对数据进行捆绑购买/销售，进一步促进大数据发挥价值。

在实际估值中，站在卖家和平台的角度还需要综合以上估值及数据采集、处理和存储的成本来形成完整的数据价值估计。

继续探索的方向

通过初步探索，我们厘清了对数据进行完整质量和价值评估的思路，但还存在以下问题需要继续探索。

1. 数据成本难以计算，其成本涉及数据采集、处理、存储等各个环节的成本，并且数据通常作为信息系统的副产物，如何核算其人力、物力成本仍是个难题。

2. 以上多个维度的评估指标并非完全相互独立，我们应该如何融合多个指标形成对数据集整体的评估结果，同时保证评估结果有很好的可解释性？

3. 面对庞大的数据集，如何减少质量评估的计算开销也是一个难题。近期被提出的一些基于深度神经网络的质量评估模型^[4,5]，虽然为部分非结构化数据的质量提供了很好的量化模型，但其在单一数据单元上近乎秒级的计算开销使其难以适用于大规模数据集。

4. 为保护卖家的所有权，买家甚至平台不能直接访问卖家所有的数据，在此情况下如何进行准确的质量和价值评估？尤其是数据的使用价值更多体现在数据被使用后带来的应用和商业价值，需要使用后才可准确获知。针对新买家的需求，我们应该如何在其购买和使用数据前合理评估数据的使用价值？

数据定价

数据的价值不等同于价格，数据应该以什么价格成交才能做到买家、卖家和平台多赢，从而保障交易市场持续健康运行，这正是数据定价需要研究的问题。目前的数据交易平台如数据堂、贵阳大数据交易所、京东万象等通常由卖家主观地发布价格，这种方法缺乏合理性，难以保障各方利益。在数据售卖过程中，我们主要考虑三种模型：(1)一个买家多个卖家，如平台采用众包机制进行数据收集；(2)一个卖家多个买家，卖家通过平台公开售卖某个数据集；(3)多个卖家和多个买家。在数据卖家/平台对数据价值进行了充分评估的情况下，标价售卖机制是一个可行的定价方案。另一个经常被采用的有效方案是拍卖，通常由买家报价，卖家根据报价情况决定将数据出售给哪位买家或流拍以优化自身利益。也可采用逆向拍卖机制，由卖家报价，买家设计机制选择卖家以优化自身利益。然而数据成本估计难、价值的动态性、复制成本低、数据的协同性、使用的排他性等为数据定价和拍卖机制的设计带来了巨大挑战。

现有研究工作

现有工作大多通过机制设计来确定最优定价，以达到买家或卖家收益的最大化。从数据价格的确定方

式上，可以分为发布价格机制和拍卖机制。发布价格又可分为固定价格和变动价格。固定价格主要取决于数据的价值估计和市场供需关系，但由于数据可反复销售，且其销售数据几乎没有边际成本，使得供需关系难以估计。变动价格则需要进一步对数据价值和市场供需的变化进行预测。当所售数据为稀缺资源（供小于求）时，可以通过拍卖来决定卖给谁和收取多少费用。精心设计的拍卖可以将社会福利（所有参与方收益的总和）或单边收益最大化。最受欢迎的拍卖类型包括英式拍卖、荷兰式拍卖和 Vickrey 拍卖等。数据拍卖的挑战之一是可能需要很长时间才能等到足够的竞标者，因此在线场景中根据等待时间可分为实时决策和延时决策两类拍卖机制。实时决策要求当新买家出价时，卖家需要立即决定是接受报价或拒绝并等待更好的报价，这个问题也称为在线搜索问题或秘书问题。一个常用的解决思路是确定一个合适的底价，并将数据出售给出价高于底价的第一个买家，而如何确定这个底价使得收益最优则是一个难题。延时决策则允许卖家设置一个时间段，在该时间段内可能有多个买家竞标，在此时间段结束后，卖家将挑选获胜者并出售数据。这类似于传统拍卖，但问题是如何设置等待的时间段。等待时间越长意味着可能有更多的买家参与，但现有的买家可能会放弃等待，同时数据价值如果是时间敏感的，那么等待中数据可能会贬值。

从另一个角度，根据卖家对买家估值信息的了解程度可以把相关机制分为两大类：一类是贝叶斯环境，即已知买家估值信息服从某个给定的分布；另一类是无先验知识环境，即没有任何买家估值信息。在贝叶斯环境中，Myerson^[9] 提出了最优拍卖机制，文献 [10] 研究并设计在买家到达顺序随机且未知的情况下，具有常量近似保证的发布价格机制。在无先验知识环境中，Goldberg^[11] 等人研究并提出一系列的线下诚实且有常量近似保证的拍卖机制。当考虑在线场景时，文献 [12] 提出通过先观察后决策的方案取得具有常量近似保证的方案。还有一系列工作探索了基于多臂赌博机问题的在线学习定价机制，从而优化收益^[13]。由于数据具有非常高的固

定成本和极低的边际成本，文献 [14] 提出了基于数据版本的定价方法，给不同版本的数据进行不同的标价以使卖家收益最大化。

在使收益最大化的同时，我们通常希望设计的机制能实现以下五个性质：

1. 激励兼容：当买家诚实报价（即出价是他对数据的真实估价）或诚实到达（在线场景下）时，每个买家都能达到最优结果。
2. 联合预防：买家不能串通起来增加收益。
3. 个人理性：每个买家都实现了非负收益，即没有人支付的费用高于他对数据的估值。
4. 预算可行性：与卖家的底价一样，每个买家都有数据预算，且不会接受高于预算的价格。
5. 计算效率高：协议可以在多项式时间内完成。

另外还有很多工作将数据质量纳入定价机制的考虑中，如文献 [13,16] 等工作探索了在预算限制下如何购买数据，使得数据多样性最大化的问题。考虑到数据中可能包含隐私信息，文献 [15] 探索了基于隐私补偿的数据定价机制，文献 [16] 探索了保护数据隐私并使数据上下文质量达到最大化的数据交易机制。

以上工作基于不同假设提出了优化收益的定价机制，但当前的定价机制大都忽略了数据价值的时间敏感性（随时间的贬值 / 增值），在数据价值时变的场景下如何定价以达到收益最大化也是重要的研究问题。

时间敏感的数据在线交易机制

面对数据价值时间敏感带来的挑战，我们设计了时间敏感的数据在线交易机制，旨在保证激励兼容，个人理性，预算可行和计算效率高的同时，使数据售卖过程中的卖家利润（近似）最大化。具体而言，我们假设数据价值随时间变化的函数 $d(t)$ 已知， $d(t)$ 和特定的待售卖的数据集相关，我们通常可以通过市场调研得到。根据卖家对买家的估值集合 V 掌握的信息多少，分几种情况设计了如下在线机制（以下均假设拍卖截止时间和买家到达速率是已知信息）。

1.V 未知：我们提出报价和到达时间诚实的机制，其按照 $d(t)$ 值的大小将时间分成多个类，每个

类中 $d(t)$ 值相差不大（如最多相差 2 倍）。随机均匀地从前 $3\log n+1$ 个类中选取某个类，然后在选中类对应的时间段执行传统的先观察后决策的方法。可证明该机制的收益竞争比是 $O(\log^2 n)$ 。

2. 买家估值信息服从已知的给定分布：我们首先提出了报价和到达时间诚实的机制，设置固定的决策价格 x ，即如果买家报价大于等于 x 则同意，并且买家支付 x 。可证明该机制的收益能够达到常量竞争比。针对数据价值下降迅速的场景，我们提出了一个报价和到达时间诚实的机制，动态地设置决策价格 $x(t)*d(t)$ 。即在时间 t ，如果买家报价大于 $x(t)*d(t)$ 则同意售出，并且买家支付 $c(t)*d(t)$ 。可证明机制的收益能够达到常量竞争比。

3. V 服从未知的给定分布：我们提出了先学习后决策的在线机制，首先在学习阶段先观察收集 m 个买家的报价，在此期间不选择获胜者，为了让买家诚实报价给予买家一定补偿。然后通过观察结果学习到报价的概率密度函数后进入决策阶段，执行买家估值信息服从已知的给定分布的算法。

待解决问题

虽然已有许多定价机制被提出，但由于数据本身的特殊性以及实际共享和交易市场的复杂性，以下问题尚待探索：

1. 数据买家之间通常存在竞争关系，买家希望能排他地购买数据以保证自己在应用市场的竞争优势。因此，在限量出售、独家垄断、所有权转让的场景下，我们应该如何确定数据销售的份数以及确定数据价格，以优化卖方的收入和买方的收益？尤其是数据具有看过即拥有、所有权模糊的特性，如何保证数据所有权的转移以及有效限制数据使用方的数量仍是一个难题。

2. 针对多方交易场景（如团购），由于多个参与方对数据的估值不同，我们应如何确定适当的折扣，保证多个参与方之间的公平性？

3. 为更好地发挥数据的协同作用，可以对多个数据集进行捆绑销售，然而捆绑后的数据可能具有更高的价值，导致数据定价变得更加复杂，同时如

何提供最优的捆绑方案，如何对多个卖家进行利益分配，如何防止买家通过组合拆分购买从而套利都是极具挑战性的问题。

总结与展望

当前数据共享和交易市场仍处于起步阶段，并朝着规范、成熟的市场不断前进。在这个进程中，对数据进行准确的质量价值评估以及合理定价是亟待解决的难题。本文围绕数据的质量、价值和价格对市场现状，相关研究，可能的解决方案以及尚待探索的开放性难题进行了讨论，希望能为相关领域的技术和科研人员带来新的灵感和思路。此外，在整个评估、共享、交易的流程中我们还必须十分重视数据的隐私和安全，否则数据共享和交易将可能成为空中楼阁和攻击的靶点。我们有理由相信，随着更多力量投入到数据共享和交易关键难题的攻克上，大量包括数据质量价值评估和定价的技术与规范将不断涌现，以形成更加健康的数据共享和交易生态，进一步释放大数据的潜在价值，有效推动经济增长。 ■



张 兰

CCF 专业会员。CCF 优秀博士学位论文奖获得者，阿里巴巴青橙奖获得者。中国科学技术大学特任教授。主要研究方向为跨域数据的深度理解、隐私保护和数据交易。zhanglan03@gmail.com



李向阳

CCF 专业会员、CCCF 编委。中国科学技术大学教授，国家千人计划专家，ACM 中国共同主席，IEEE Fellow。主要研究方向为大数据的共享交易和隐私保护等。xiangyang.li@gmail.com



李安然

中国科学技术大学博士研究生。主要研究方向为图数据深度学习、数据交易、数据质量分析。
anranLi@mail.ustc.edu.cn

其他作者：薛爽爽

参考文献

- [1] Li X, Qian J, Wang X. Can China lead the development of data trading and sharing markets? [J]. *Communications of the ACM*, 2018, 61(11):50-51.
- [2] 李向阳, 张兰, 韩风, 等. 大数据共享及交易中的机遇和挑战 [J]. 中国计算机学会通讯, 2019, 15(1):43-51.
- [3] Rezaee Z. *Intangible asset valuation* [M]. Financial Services Firms: Governance, Regulations, Valuations, Mergers, and Acquisitions(Third Edition). 2011:331-344.
- [4] Putra J. W. G. and Tokunaga T. Evaluating text coherence based on semantic similarity graph[C]//*Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, 2017:76-85.,
- [5] Kang L, Ye P, Li Y, and Doermann D. Convolutional neural networks for no-reference image quality assessment[C]//*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014.
- [6] Loizou P C. *Speech quality assessment* [M]. Multimedia Analysis, Processing and Communications. Springer, 2011: 623-654.
- [7] Gil Y and Artz D. Towards content trust of web resources[J]. *Web Semantics: Science, Services and Agents on the World Wide Web*. 2007,5(4):227-239.
- [8] Batini C, Cappiello C, Francalanci C, and Maurino A. Methodologies for data quality assessment and improvement[J]. *ACM Computing Surveys*, 2009, 41(3):16.
- [9] Myerson Roger B. Optimal auction design[J]. *Mathematics of Operations Research*. 1981,6(1):58-73.
- [10] José Correa, Patricio Foncea, Ruben Hoeksma, et al. Posted price mechanisms for a random stream of customers[C]//*ACM Conference on Economics and Computation*. ACM, 2017:169-186.
- [11] Goldberg Andrew V, Hartline Jason D, and Wright Andrew. Competitive auctions and digital goods[C]//*The Twelfth ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, 2001:735-744.
- [12] Blum Avrim and Hartline Jason D. Near-optimal online auctions[C]//*The Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.
- [13] Zheng Z, Peng Y, Wu F, et al. An online pricing mechanism for mobile crowdsensing data markets[C]//*Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc)*, 2017.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

数据交换

关键词：数据资源 数据交换

左 淞
谷歌研究院

导论

随着计算资源与处理技术的发展，潜藏在海量数据中的复杂结构信息得以被人们发掘并加以利用。大数据时代应运而生，数据资源更是被誉为新时代的石油。同石油一样，数据也是一种稀缺资源，其价值与稀缺性必然会在人类社会中催生出围绕数据资源的经济活动。

数据资源本身及其价值产生方式的特殊性，使得其在经济活动中的价值与石油等传统资源存在着本质区别。这些区别也使得针对传统资源的交换机制并不能直接应用到数据资源的交换中。

在经典的交换模型中，参与者带着各自的物品加入到交换机制中，在不涉及金钱交换的情况下，通过换取其他参与者带来的物品以获得更好的分配组合。例如房屋分配^[1,2,15]以及器官移植交换^[3,7,12-14,16]等。

如果直接将这些模型应用到数据交换中，则会遇到一些问题。例如当交换发生时，物品的原拥有者将失去该物品，而在数据资源问题中则并非如此^[4,8-11]。在另一方面，随着某个具体的数据资源被更多人拥有，其稀缺性的下降也给原来的拥有者带来负外部性。

这些数据交换问题中的特性使得交换中参与者的偏好与策略变得与传统交换问题不同。进而使得交换机制下的博弈问题发生改变。事实上，在一些较为直观的交换机制中，判定纯策略纳什均衡¹的

存在性也是 NP 难 (NP-hard) 问题^[6]。

数据资源及其价值

数据资源的价值来源是其中包含的信息，这些信息被有效地发掘并利用到人类社会活动中时，就产生了价值。从经济学理论的角度来讲，这些价值的根本是对稀缺资源更加高效的分配。

直接提高资源分配的高效性 (efficiency)。一个典型的例子是通过用户画像等精准定位技术提高市场营销的效率，将有用的信息告诉需要的人。相对于传统的电视或者街头广告等形式，基于精准定位的在线商业推广可以极大地提高针对目标人群的投放效率，同时也更好地避免了过量无关信息骚扰导致的用户反感。这种商业模式在过去的十年中造就了一批世界顶级的巨头企业。例如 Google、Facebook、淘宝、百度、字节跳动等企业，互联网流量分发都是其重要收入来源。尽管其最初依赖互联网的便利性带来了用户的增长，但当用户接近饱和之后更多的是依靠流量分配效率的提升，这种效率提升正是来源于从数据资源中挖掘得到的有用信息。

通过对某种生产项目效率的提高，减少这种生产项目对资源的消耗，从而间接提高各生产项目间的整体资源分配效率。这方面代表性的例子是人工智能技术的应用。近年来，随着数据量的增长及处理能力的提高，基于人工神经网络的人

¹ 纳什均衡指博弈中这样的局面，对于每个参与者来说，只要其他人不改变策略，他就无法改善自己的状况。

工智能技术的实际效果提升迅速，并在许多方面达到甚至超越了人类的水平，如在计算机视觉，机器翻译，一些病理诊断，棋牌类博弈，甚至新天体发现等科研问题上。随着这些技术的应用，部分劳动力资源的消耗可能会减少，从而促进劳动市场资源分配的优化，产生社会价值。

数据资源的稀缺性

数据在人的生产生活中并不稀缺，缺少的是被详尽地收集并记录下来的数据资源。“大数据”一词的活跃时间不过五年左右，而在更早之前，由于缺乏有效的挖掘与利用信息的技术手段，以可利用资源的形式聚集起来的数据资源非常有限。事实上，能够比较高效地处理大规模数据的成熟手段，源自基于 MapReduce 架构的分布式计算机集群^[5]，后来也被抽象为“云”(cloud)这一概念。

大规模数据资源积累的历史较短，其有效渠道也相当稀缺。一方面，敏感用户隐私数据的收集受到法律的约束及监管，如 2017 年 6 月 1 日起生效的《中华人民共和国网络安全法》，2018 年 5 月起在欧盟实施的《通用数据保护条例》(GDPR) 等。另一方面，用户使用一些服务的行为数据往往只有相关的服务提供者能够获取，甚至在一些互联网公司，不同部门之间也存在着内部数据壁垒。

数据资源的特殊性

根据最基本的经济学原理，数据的价值及其稀缺性必然会在人类社会中催生出围绕这种资源的经

济活动。而由于数据资源本身的一些特殊性，其交易属性与传统的商品有着明显的不同。

数据几乎可以被零成本地完美复制。与传统商品交易的从“甲拥有乙没有”到“甲没有乙拥有”不同，数据资源的交易则是从“甲拥有乙没有”到“甲乙均拥有”的状态（如图 1）。事实上，由于甲很难证明自己不再拥有这些数据，传统意义上的将数据商品由甲完全地交割给乙是无法实现的。

数据稀缺性下降产生的负外部性。因为交易之后甲乙均拥有了同样一份数据，原则上二者可以同样地使用这份数据并从中挖掘出相同的信息。如此一来，双方在同一目的上的潜在竞争将会导致该数据资源在双方共同占有下，对一方（如甲）的价值低于其仅由一方（如甲）占有时的价值。同理，当该数据资源被更多人共同占有时，其价值也将随着分享的人数增加而降低。

传统商品的市场和交易模型无法直接用来描述针对数据资源的市场和交易行为。因此，我们需要一个经过修改的新模型以描述上述两种特殊性，并重新研究其中的市场行为与均衡。

一个简单的模型

在实际中，较为简单的数据交换模式是经由一个平台自由交换数据。用户首先在平台上注册会员，并提供一个具有一定价值的数据集，经过平台审核后该数据集将在平台上进行共享。会员也可以在平台上免费获取共享的其他数据集。如果忽略掉其中的少量费用，我们可以用一个简单的无金钱模型来研究在该平台下进行数据交换的博弈与均衡。

在这个简单的模型中，每个人拥有一个独有的数据集，并且可以决定是否参与到数据共享之中。如果要参与数据共享，他必须将他所独有的数据集在参与者中共享，而他则可以获取一些由其他参与者共享的数据。我们假设每个人对每个数据集有价值估计，其参与到数据共享中获得的总效用，则是将从其他人分享的数据集中获取的价值减去将自己的数据集分享给他人带来的损失。

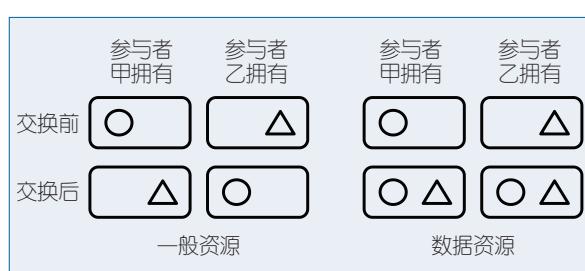


图 1 不同资源交换对比

为了简化分析，我们考虑一种比较特殊的线性负外部性，即一个数据集对每个人的价值均随着使用该数据集人数的增加而严格减少。每个人能够采取的策略则是选择不加入共享或者共享并选择使用所有数据集（处于共享状态）中的一部分。

纯策略均衡

在这个简单的模型下，我们通过分析其均衡来了解该共享平台的性质。由于实际中每个参与者所采取的策略及其共享状态均可被其他的参与者观察到，研究纯策略均衡的性质能够更加合理地反映现实中可能出现的博弈均衡。

问题一：这样的纯策略均衡是否一定存在呢？

首先，该博弈始终有一个平凡的均衡，即所有人都选择不参加。因为在这种情况下，每个人仅改变自身的策略并不能使其获得任何新的数据。值得注意的是，除了这样的平凡均衡，非平凡的均衡并不总是存在的。考虑如下三个参与者（甲乙丙）的例子：

1. 甲因为对自己的数据估值很低，而对其他人的估值很高，总是愿意加入共享；
2. 乙对自己的数据以及甲的数据很看重，愿意与甲两人分享，然而一旦有第三人加入，乙则会因为分享自己的数据带来的损失超过获取甲的数据带来的收益而选择停止共享；
3. 丙对自己的数据比较看重，但是认为乙的数据价值远高于自己的数据，因此即使三人分享也愿意加入以获得乙的数据，然而丙认为甲的数据没有价值，所以一旦乙没有参与共享则会选择停止共享。

表1 参与者数据分析

参与者	策略非均衡者	改变策略动机
不包含甲	甲	获取乙或丙的数据
甲	乙	获取甲的数据
甲乙	丙	获取乙的数据
甲乙丙	乙	获取甲数据的收益小于分享自己数据的损失
甲丙	丙	获取甲数据的收益小于分享自己数据的损失

经过简单的分析（见表1），可以发现除了所有的人都不参与共享以外，任何一种参与共享的组合都不构成均衡：甲没有参与的任何组合中，甲都会愿意改变策略参加共享；如果只有甲参与了共享，那么乙则愿意改变策略参加共享；如果甲乙均参加了共享，那么丙也愿意改变策略参加共享；如果甲乙丙均参加了共享，那么乙会改变策略停止共享；如果只有甲丙参加了共享，那么丙则会改变策略选择停止共享。

问题二：对于给定的博弈，我们是否能够快速地确定其是否存在非平凡均衡，并找出这样的均衡呢？

我们可以证明判定是否存在非平凡均衡是一个NP难的问题^[6]。该结论可以通过将任意3SAT问题归约为判定是否存在非平凡均衡的问题，其构造的思想则可以基于上述不存在非平凡均衡的例子。值得注意的是，即使我们在原问题中限制每个人最多只能从平台上获取两个新数据集，该问题仍然是NP难的。但在适当的对称性假设和单一选择的限制下，在多项式时间内找出非平凡均衡的算法是可行的。

结语

关键生产资源的分配始终都是人类经济活动中的核心问题。这里我们仅仅基于一个最简单的模型对一个全新的领域进行了小小的展望。希望能够抛砖引玉，提供一点新的视角。 ■



左 淼

谷歌研究院博士，曾获Google PhD Fellowship（谷歌博士奖研金）。主要研究方向为拍卖理论、机制设计、计算经济学、博弈论等。songzuo.z@gmail.com

参考文献

- [1] Abdulkadiroğlu A, Sönmez T. Random serial dictatorship and the core from random endowments in house allocation problems[J]. *Econometrica*, 1998, 66(3):689-701.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

个人数据交易：从保护到定价

关键词：个人数据市场 数据服务 安全隐私 定价机制

牛超越 郑臻哲 吴帆 等
上海交通大学

个人数据流通和交易现状

在大数据驱动经济发展的今天，个人数据已经日趋商品化。国内外知名的互联网公司往往通过提供免费的在线服务以获取个人数据，许多大数据创业公司通过支付用户酬劳来获取对数据的使用权限。然而，当互联网用户逐渐意识到伴随数据分享而来的隐私泄露等风险时，将拒绝在自己的敏感信息没有得到全面保护、隐私泄露没有得到合理补偿的情况下提供个人数据。政府部门和企业作为个人数据资源的主要采集者和拥有者，也存在用户隐私保护和商业机密隐藏等安全需求，他们大都只在内部分析和使用用户数据。海量的个人数据缺乏开放和流通，形成了大量的信息孤岛，严重地抑制了市场对数据的需求，成为大数据发展的瓶颈，亟须安全可靠的数据交易平台促进个人数据资源开放、推动数据应用和释放数据价值。

为了促进个人数据流通，许多数据代理商纷纷出现，在数据贡献者和数据消费者之间搭建桥梁：一方面，通过金钱性的补偿来激励数据贡献者分享数据；另一方面，为数据消费者提供数据服务并收取一定的费用。根据美国联邦贸易委员会(Federal Trade Commission)在2014年5月发表的关于九个具有代表性数据市场的调查表明：总部位于美国阿肯萨斯州的安客诚公司(Acxiom)作为最大的数据代理商从全球约7亿用户处采集个人数据，并为全球顶尖的企业提供基于数据智能分析的商业解决方案^[1]。同年8月，美国哥伦比亚广播公司新闻部门的《60

分钟》节目(CBS News 60 Minutes)对此提出质疑：数据代理商从用户的个人数据中获取暴利，却没有对用户的隐私泄露进行补偿^[2]。围绕数据交易市场构建的相关话题也引起了我国政府相关部门的高度重视，并引发了新闻媒体的广泛讨论。2015年9月5日，《国务院关于印发促进大数据发展行动纲要的通知》正式发布。该行动纲要的核心是推动各部门、各地区、各行业、各领域的数据资源共享开放，并明确提出要引导培育大数据交易市场，建立健全数据资源交易机制和定价机制^[3]。工业和信息化部于2017年1月发布的《大数据产业发展规划(2016—2020年)》进一步指出要研制数据资源分类、开放共享、交易、标识、统计、产品评价、数据能力、数据安全等国家通用标准^[4]。在国家政策的积极鼓励以及地方政府和产业界的带动下，数据交易从概念逐步落地，贵州、武汉等地的数据交易平台先后投入运营，并在数据定价、交易模式、交易标准等方面进行了有益的探索。2018年5月，国家信息中心的大数据研究专栏发表了3篇文章评析我国大数据交易的发展现状和面临的困难，呼吁制定国家层面的法律规范^[5]。

个人数据市场框架

个人数据市场^[6~8]主要有三种类型的参与者：数据贡献者、数据代理商以及数据消费者，框架如图1所示。该框架主要包括数据采集层和数据交易层。

在数据采集层，数据代理商从数据贡献者处采

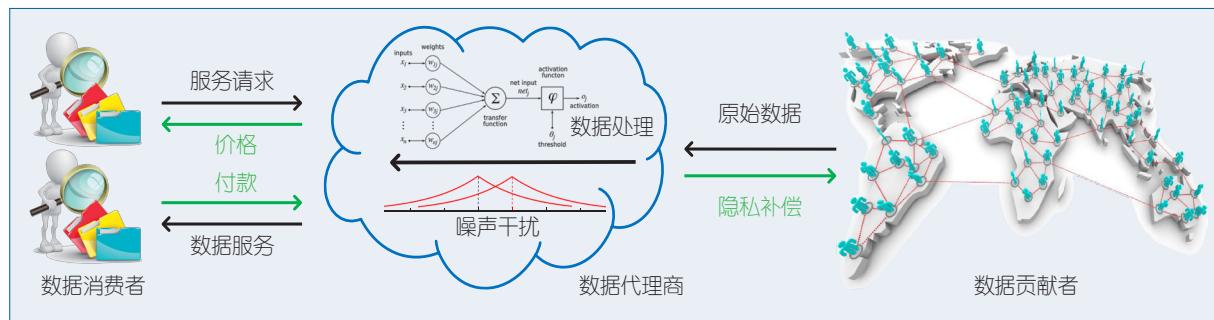


图1 基于干扰型数据服务的个人数据市场框架

集大量的个人数据，例如社交媒体数据、运动轨迹、医疗记录、住宅能源消耗量、用户评分等。由于现实生活中存在社交、行为、基因等多种多样的交互活动，实际数据之间往往存在复杂的关联性，例如，互为好友的两个用户的活动轨迹在时间和空间维度上具有很强的一致性。此外，不同类型的数据具有一些其他的特征，例如，一些用于决策的数据具有很高的时效性；时空感知数据具有不完整、不精确、易错误等特性。因此，实际可行的数据交易框架应该将数据的具体特征考虑在内。

在数据交易层，数据代理商倾向于交易数据服务而非原始数据。这里的数据服务是数据处理产生的结果。相比于交易原始数据，提供数据服务主要有以下优势：对数据贡献者来说，数据服务更能保护隐私；对数据代理商来说，原始数据的所有权和版权难以界定和管理，同时，增值的数据服务的市场需求更大，产生的收益更高；对数据消费者来说，数据服务更能直观深入地刻画整个数据集潜在的特征与规律，充分发挥大数据的实用性。此外，每个数据消费者可以自定义服务请求。值得注意的是，除了具体的数据处理方法，服务请求还包括在返回的数据处理结果中添加噪声的等级，例如噪声服从分布的方差。数据代理商采用随机化机制回答服务请求：返回结果的期望是数据处理准确的结果，而方差不能超过所指定的方差。相比于普通的数据服务，干扰型数据服务允许数据消费者选择适合自己精确度需求的数据服务并支付相应的价格。

根据具体的服务请求，数据代理商一方面向数据消费者收取特定的费用，另一方面需要补偿数据

贡献者的隐私泄露。服务请求中的噪声干扰等级越高，返回的数据服务越不准确，数据消费者需要支付的费用也越低，隐私泄露程度越低，数据贡献者的隐私补偿也应该越少。鉴于整个数据市场框架需要满足收支平衡，即数据代理商的收益必须大于等于零，这也意味着数据服务的价格要大于等于隐私补偿的总和。

关键研究问题

安全可信的数据交易环境

数据市场的安全性研究主要针对私密保护和可验证性这两个有机关联的问题。首先，数据贡献者在数据市场中不希望泄露自己显式的身份标识，例如身份证号、手机号等，即身份私密性；其次，数据贡献者和数据消费者都不希望暴露自己敏感的数据内容，例如运动轨迹、医疗记录等，即数据私密性；再次，数据代理商不希望泄露自己的数据处理模型，例如支持向量机中的支持向量、神经网络中的权值矩阵等，即模型私密性。在保护各参与方敏感信息的前提下，数据交易还有着可验证性的安全需求。数据消费者不仅需要验证数据来源的真实性，还需要验证数据处理结果的正确性与完整性。

研究现状

个人数据有广泛的应用前景，能够提供高价值、高质量的信息，但是如果直接将个人数据用来交易，将存在隐私泄露等问题。近几年，研究者逐渐将目光聚焦到数据市场的安全机制设计。在文献[9]提

出的 DataLawyer 系统中，数据代理商能够显式地制定数据使用规则，并且能够在数据消费者使用数据的时候自动检测数据使用是否满足相应的规则，以确保数据不被非法使用。典型的数据使用规则包括数据溯源、限制查询频率、禁止多元数据聚合等。

中国科学技术大学教授李向阳的团队考虑了不可信的数据消费者二次贩卖数据集的问题^[10]，并将此问题转化成集合相似度的比较问题。他们考虑了文本数据、视频图像数据和图表数据等多种类型的数据。最近，李向阳教授团队针对语音数据提出了保护隐私和可用性的数据发布方案^[11]。他们针对图片数据提出了基于群智感知的大规模、高质量的数据采集方案，并考虑了所有权和隐私等安全问题^[12]。

存在问题：缺少针对数据市场三方模型全面立体的保护机制

已有的相关研究工作主要考虑了数据市场中的单个环节，也相应采用了密码学系统中经典的两方模型，但没有全面考虑数据市场新颖的三方模型，因此很难在现实的数据市场中真正得到应用。数据市场三方模型使得设计安全保护机制有了新的挑战，主要体现在数据贡献者有保护个人隐私的需求，数据代理商有保护数据处理模型私密性的需求，数据消费者有验证数据采集与处理真实性的需求。现有的密码学工具，例如数字签名机制，是在泄露身份隐私的情况下保证数据采集的真实性验证。此外，数据处理的真实性验证与经典的外包计算场景中的可验证性计算^[13~20]有着本质区别，即数据消费者作为验证者并不知道原始数据集与外包函数。

考虑到私密保护与可验证性的内在矛盾性，如何保证数据来源的真实性且维持身份隐私与数据私密性，如何保证数据处理结果的正确性与完整性，且不破坏数据私密性与模型私密性，这两个都是极具挑战性的研究问题。目前在密码学领域出现的理论工作大都只保证其中的一个性质。此外，所假设的系统模型与新颖的数据市场三方模型有着本质区别。因此，我们需要充分考虑数据贡献者、数据代理商、数据消费者的私密保护与可验证性需求，研究实际可用的数据市场保护机制，提供全方位安全

可信的数据交易环境。

初步探索

我们从数据和模型的私密性、数据来源及数据处理结果的可验证性两方面研究具有私密可保护的可验证数据交易机制。

由于需要对数据代理商隐藏数据贡献者的原始数据或数据消费者的输入数据，同时还需要保证数据代理商能够高效地处理原始数据和服务请求，我们无法直接使用传统的对称加密算法或非对称加密算法。对于数据私密性的保护，同态加密技术是一种值得考虑的方法，包括支持在密文上同时进行加和乘运算的全同态加密系统与支持一些指定运算的部分同态加密系统。同时，同态加密技术也可以应用到数据处理模型参数的保护，而不损失数据处理模型在数据消费者这一方的可用性。

数据消费者需要验证数据来源的真实性，这就要求数据贡献者在密文上进行签名。值得注意的是，数据来源的真实性验证类似于数字签名机制中的不可抵赖性，而不是数据的真实性与完整性，因为数据消费者是作为第三方而不是数据接收方来验证数据发送方的合法性。数据消费者验证数据处理结果的真实性与完整性，最为简单直接的方式就是利用同态性质重新计算，并且检查两次结果的一致性。当然在这种朴素的设计方案中，数据消费者需要额外花费昂贵的计算开销，这正是我们的结果验证协议需要避免的地方。

待解决问题

同态加密算法与数字签名算法的结合为敏感数据保护提供了基本的技术思路，但仍存在四方面的问题。第一，数字签名算法采用的是顺序验证的方式，需要花费较大的计算开销，传输数字签名和维护数字证书也需要花费较大的通讯开销。因此，该环节可能会成为大规模数据市场的瓶颈。第二，已有的签名算法大都把签名者的身份标识当作公共参数，而在实际的数据市场中，数据贡献者作为签名者希望保护自己的身份标识。但是，如果隐藏了所有的身份标识，数据市场中的违法数据贡献者就难以被发现，即需要解决身份隐私与可追溯性之间的

矛盾。第三，完全同态加密协议虽然可以支持密文上多种类型的计算，但其较高的计算开销仍然无法很好地适应大规模数据市场的需求。第四，数据处理结果正确性与完整性的证明通常涉及数据消费者的输入数据与数据处理模型的参数，这与保护数据私密性与模型私密性的目标相违背。同时，数据服务请求的大规模性与数据处理模型的复杂性对于验证协议的设计提出很高的性能要求。

自底向上的定价机制

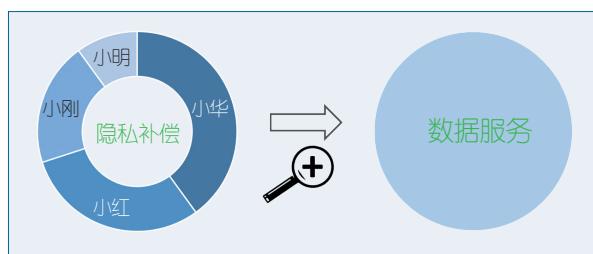


图 2 自底向上的定价机制设计

在数据交易场景下，数据消费者需要获得数据处理的结果，数据代理商则需要衡量每次所提供的数据服务对于每个数据贡献者的隐私泄露程度，并对他们进行合理的隐私补偿。从这个角度来看，隐私补偿机制可以看作传统的激励机制在个人数据采集场景下的变种。如果将隐私补偿机制放入整个数据市场的定价框架中，并采用自底向上的设计思路（图 2），即底层的隐私补偿总和决定上层的数据服务价格。隐私补偿机制是数据市场定价机制的核心与基石。

研究现状

数据市场需要高效的数据采集机制来为数据代理商不断补充优质、海量的数据资源，从而保证数据市场的持续繁荣。众包机制被认为是大规模数据采集的一种有效方法，其核心问题是如何设计激励机制以提高用户参与度。Lee 等人设计了基于动态定价的逆向拍卖机制^[21]，该机制的优化目标是使众包平台数据采集的花费最小化，并保证一定的用户参与量。Jaimes 等人考虑了数据贡献者的地理位置信息，并将数据采集建模成存在预算限制的覆盖最

大化问题^[22]。以上两个工作并没有考虑众包平台中可能存在的操纵策略。Yang 等人将用户的策略行为建模成两种不同的博弈模型：以平台为中心的模型和以采集用户为中心的模型，并分别采用斯塔克尔伯格 (Stackelberg) 博弈和逆向拍卖来设计数据采集机制^[23]。清华大学的杨铮等人提出了三种在线激励机制，以处理数据采集过程中数据贡献者随机出现的情况^[24]。针对个人数据采集的场景，亚利桑那州立大学 Wang 等人研究了在数据代理商不可信的情况下如何通过激励机制购买添加噪声干扰的隐私数据，并建立了博弈模型来衡量隐私的价值^[25]。

由于采集数据的质量参差不齐，众包平台还需要设计评估数据质量的管理方案以引导数据贡献者提供高质量的数据。Liu 等人借鉴在线学习的思想来提高众包平台采集数据的质量^[26]。Karger 等人利用推断算法来检测数据冗余^[27]。此外，众包平台中衡量数据质量最为棘手的问题是真实数据的缺失，即众包平台不仅需要衡量数据贡献者的数据质量，还需要预测真实数据。数据挖掘领域中真值发现 (truth discovery) 框架为处理该问题提供了行之有效的方案^[28]，其核心思想类似于期望最大化算法。然而这些工作并没有将数据质量衡量与酬劳机制联系起来，无法从本质上激励用户贡献高质量的数据。Peng^[29] 和 Jin^[30] 等人考虑了基于数据质量的酬劳激励机制设计。Jin 等人进一步将真值发现拓展到策略博弈环境^[31]。

上述激励机制研究工作的设计目标主要集中在社会效益最大化和数据采集酬劳开销最小化两方面。哈佛大学陈怡玲 (Yiling Chen) 研究组系统地研究了在策略博弈环境下机器学习任务导向型的数据采集方案^[32-34]。文献 [35] 介绍了如何为线性回归模型设计真实可信的数据采集机制。文献 [36] 将该采集机制拓展到更普适的回归模型。文献 [37] 将线性回归建模成非合作博弈模型，并充分考虑用户在数据采集过程中的隐私。文献 [38] 考虑了激励相容条件下的回归学习框架。

存在问题：未考虑数据的实际特征，隐私泄露的量化不准确

初步探索

设计合理的隐私补偿机制的主要挑战在于如何准确地衡量每个数据贡献者的隐私泄露程度，但当前学术界和工业界与个人数据相关工作的主要出发点与落脚点是保护隐私，例如，谷歌^[39]、苹果^[40]、微软^[41]等公司利用差分隐私框架来保护用户隐私。从统计科学角度来看，差分隐私的目标是尽可能多地挖掘关于整体数据集的规律，同时尽可能少地泄露个人的信息。我们近期的研究工作^[8]发现：量化隐私泄露本质上是保护隐私的逆过程。因此，我们可以利用隐私保护机制中的基本原理和准则实现量化隐私泄露的目标。个体隐私泄露可以定义为有无某个数据贡献者的数据对于数据处理结果分布的影响。这里的“有无某个数据贡献者的数据”代表着差分隐私框架下的一对相邻数据库。

在数据关联性方面，我们采用广义的差分隐私框架——河豚隐私 (Pufferfish privacy)^[42,43]。河豚隐私主要通过引入数据分布这一参数来更好地保护关联型数据的隐私。这里的数据分布主要用来形式化数据关联性，例如，社交网络数据中的关联性可以用贝叶斯网络来刻画，而时间序列数据中的关联性可以用马尔可夫链来表达。

待解决问题

差分隐私和河豚隐私为我们定义隐私泄露提供了基本的框架，但距离实际的隐私泄露衡量以及后续补偿机制的设计还有以下关键问题有待解决。第一，计算可行性和计算有效性。如果直接利用上述定义精确地计算隐私泄露，数据代理商需要考虑所有可能相邻数据库的实例，这在大规模数据集上是计算不可行的。第二，从隐私泄露的衡量跨越到隐私补偿机制的设计，需要考虑数据贡献者不同的隐私策略，实用可行的隐私补偿机制需要考虑补偿方案的多样性和可满足性。第三，隐私补偿方案的设计还需要考虑公平性。文献[44]提出原始的公平性是指：如果数据处理没有涉及某个数据贡献者的数据，那么该数据贡献者获得的补偿为零。我们需要论证该公平性的定义是否适用于数据之间存在关联性的情况，如果不适用，应该如何定义广义的公平

性。第四，数据关联性更新算法的高效性。随着时间或地点的变化，数据之间的关联性也会随之改变，重新计算的成本较高，因此需要提出增量式数据关联性的度量方法。第五，隐私补偿的本身也会造成数据贡献者隐私的泄露，需要设计相应的保护机制。

自顶向下的定价机制

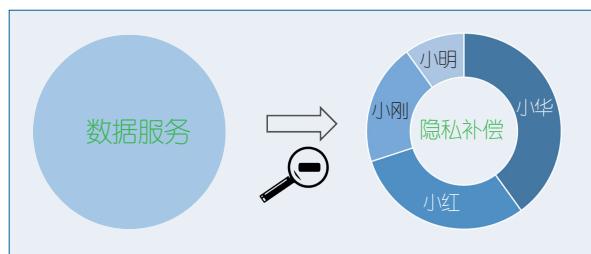


图3 自顶向下的定价机制设计

数据代理商还可以采用自顶向下的数据服务定价机制（见图3）。数据代理商匀出部分数据服务收入作为隐私补偿的预算，并按照隐私泄露的大小决定每个数据贡献者的份额，而数据贡献者不需要主动参与到隐私补偿的过程。如何对干扰型数据服务定价是自顶向下定价机制的核心。

研究现状

数据定价已经成为计算机领域和经济学领域热门的研究话题。近年来，数据库领域涌现出许多研究关系型数据的定价工作。华盛顿大学 Dan Suciu 教授领导的研究组是这个方向的开拓者。在他们最早的数据定价文章中，Balazinska 等人展望了数据交易市场的前景，并且提炼出数据交易这个方向可能的研究问题^[45]。Koutris 等人^[46,47]提出了基于查询的数据定价 (query-based data pricing) 框架，并指出数据定价中两个重要的性质：无套利性 (arbitrage freeness) 和无折扣性 (discount freeness)。文献 [48] 提出了无套利的、适用于任何查询方式的定价函数。文献 [49] 提出了针对动态数据的定价方案。最近，Deep 等人在依赖于结果 (answer dependent) 和独立于实例 (instance independent) 两种不同设定下刻画了无套利定价函数的特征^[50]。基于该理论工作，他们还实现了支持大规模关系查询定价的原型系统^[51]。

上述数据库领域的数据定价相关工作关注的大多是结构化、关系型的通用数据。个人数据已经被很多数据代理商采集和分析，并且售卖给其他数据消费者来进行精准的市场营销^[52-54]。美国纽约大学的 Laudon 教授早在 1996 年就从经济学角度构想了可以交易个人数据的全国性信息市场^[55]，但是关于个人数据定价的严格理论研究则出现在 2011 年：雅虎研究院的 Ghosh 与微软研究院的 Roth 把差分隐私作为量化隐私泄露程度的指标，并提出以拍卖的形式交易隐私数据^[56]。他们主要考虑的应用场景是单次的计数查询。Li 等人的后续研究工作，通过引入无套利的概念将应用场景拓展到多次的线性查询^[44]。Dandekar 等人讨论了隐私拍卖在推荐系统中线性预测函数以及预算限制下的应用^[57]。在近期《美国计算机学会通讯》(CACM) 的两篇文章^[58,59]中，Roth 和 Li 等人总结并展望了个人数据的定价问题。

存在问题：数据服务形式单一，缺少强健的干扰型数据服务定价机制

虽然数据定价的相关研究成果已经有很多，但大都仅适用于数据库查询，而不能直接应用于现实生活中的数据服务。相比于数据库查询，数据服务中所涉及的数据处理方法往往呈现出更加复杂多样的数学形式。例如，在聚合统计场景中，不同的统计方法涉及的算子不尽相同：加权求和涉及的算子是一次多项式，高斯分布拟合中涉及的算子是二次多项式，而度分布中涉及的算子是非线性的比较操作。因此即使为同种类型的数据服务制定统一的定价策略也颇具挑战性。此外，大部分研究工作主要关注的是一般通用数据而非个人的私密数据，因此未考虑在添加不同程度干扰噪声的情况下如何定价。

初步探索

数据服务的定价策略规避套利机会的核心挑战在于解决数据服务之间的相互决定关系，而此问题类似于数据库领域中已有的查询 / 视图的可回答性问题，例如，判断某个 SQL 查询能否通过其他的查询组合进行回答。当然，数据服务呈现出更为复杂多样的形式，之间的决定关系既可能是简单的线性关系，也可能是复杂的非线性关系。关于线性的决

定关系，已有研究工作^[54,69]发现了其与线性代数中的半范数 (semi-norm) 具有结构上的相似性以及理论上的等价性。而对于非线性的决定关系，主要有两种不同的研究思路：第一种是通过引入界面数据库的方式，简化数据处理的中间过程，只考虑数据处理的最外层的数学操作。例如，我们将常见的聚合统计建模成“点积”操作的形式，并建立线性的决定关系^[16]。第二种是通过研究数据处理的计算公式，发现与其具有相似结构的数学概念，最后建立两者之间的等价关系。

关于数据服务定价函数中涉及噪声干扰的部分，我们首先需要确定噪声所服从的概率分布，还需要准确地定义噪声干扰的等级。对于同种数据服务，噪声干扰等级越高，数据服务的精确度越低，所对应的价格也应该越低，即数据服务的价格与噪声干扰的等级之间存在负相关的关系。现有的理论推导结果表明：当噪声方差在数据服务定价函数中相对独立时，数据服务无套利的定价函数随着噪声方差的递减速度不能超过线性^[8,44,59]。

完整的干扰型数据服务的定价函数需要有机地整合数据处理模型以及噪声干扰等级两个部分，并且确定两者之间是否会相互影响，现有的研究工作都是显示地或隐式地假设两部分相对对立。

待解决问题

干扰型的聚合统计定价为我们解决一般性的数据服务定价提供了基本思路，但依然存在问题。第一，从数学上解决非线性函数的决定关系，并将其归约到现有的问题上，例如集合覆盖、网络流问题等。第二，在问题一的基础上，针对具体的、相对复杂的数据处理算法进行合理的建模并设计出鲁棒性的定价机制。第三，如果噪声干扰等级在数据服务定价函数中不是相对独立的，那么如何保证无套利的性质？最后，我们需要权衡无套利的经济学性质所带来的“利”与“弊”。其中，“利”针对贪婪的数据消费者，他们需要拥有较大的算力，花费较大的开销去发现套利的机会，并且他们的攻击开销不能超过所获取的利润。从本质上来说，无套利的定价函数意味着套利攻击的计算不可行性；“弊”针

对数据代理商，他们需要从理论上保证无套利的性质，因此数据服务的定价函数必须保有严格的数学性质，例如次可加性 (subadditivity) 等，这也意味着可供选择的定价函数的种类非常有限。

总结

个人数据交易是目前数据库、数据挖掘、机器学习、网络与信息安全、经济学等多个研究领域中热门的、学科交叉的话题，并得到了学术界、工业界以及政府相关部门的高度重视。在个人数据市场框架中起到基石作用的保护机制和定价机制的相关研究方兴未艾。我们对安全可信数据交易环境的搭建、隐私泄露量化和补偿机制的设计、干扰型数据服务定价机制的设计等三个主要方面进行了初步的探索，希望能激发相关领域研究者的探索兴趣，实现大规模个人数据交易的健康化和产业化，充分释放大数据的潜力，助力国民经济的蓬勃发展。 ■



牛超越

上海交通大学计算机科学与工程系博士生。主要研究方向为数据隐私和可验证性计算。

rvince@sjtu.edu.cn



郑臻哲

CCF 专业会员。2018 CCF 优秀博士学位论文奖获得者。上海交通大学计算机科学与工程系博士后。主要研究方向为算法博弈论、云计算、无线网络。

zhengzhenzhe220@gmail.com



吴帆

CCF 专业会员。上海交通大学计算机科学与工程系教授。主要研究方向为网络经济学、无线网络、移动计算、隐私安全。

fwu@cs.sjtu.edu.cn

其他作者：陈贵海

参考文献

- [1] Federal Trade Commission (FTC). Data brokers: A call for transparency and accountability[OL]. <https://www.ftc.gov/reports/data-brokers-calltransparency-accountability-report-federal-trade-commission-may-2014>.
- [2] The data brokers: Selling your personal information[OL]. (2014-03-09). <https://www.cbsnews.com/news/the-data-brokers-selling-your-personal-information/>.
- [3] 国务院关于印发促进大数据发展行动纲要的通知 [OL].(2015-09-05).http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
- [4] 工业和信息化部关于印发大数据产业发展规划（2016—2020年）的通知 [OL]. (2017-01-17). <http://www.miit.gov.cn/n1146295/n1652858/n1652930/n3757016/c5464999/content.html>, 2017.
- [5] 国家信息中心大数据研究 [OL].(2018). <http://www.sic.gov.cn/Column/551/0.htm>.
- [6] Niu C, Zheng Z, Wu F, et al. Trading data in good faith: Integrating truthfulness and privacy preservation in data markets[C]//Proc. of ICDE. IEEE, 2017: 223-226.
- [7] Niu C, Zheng Z, Wu F, et al. Achieving data truthfulness and privacy preservation in data markets[J].IEEE Transactions on Knowledge and Data Engineering, 2019, 31(1): 105-119.
- [8] Niu C, Zheng Z, Wu F, et al. Unlocking the value of privacy: Trading aggregate statistics over private correlated data[C]//Proc. of KDD. ACM, 2018: 2031-2040.
- [9] Upadhyaya P, Balazinska M, Suciu D. Automatic enforcement of data use policies with datalawyer[C]// Proc. of SIGMOD. ACM, 2015: 213-225.
- [10]Jung T, Li X, Huang W, et al. Accountrade: Accountable protocols for big data trading against dishonest consumers[C]//Proc. of INFOCOM. ACM, 2017: 213-225.
- [11]Qian J, Han F, Hou J, et al. Towards Privacy-Preserving Speech Data Publishing[C]//Proc. of INFOCOM. ACM, 2018: 1079-1087.
- [12] Zhang L, Li Y, Xiao X, et al. CrowdBuy: privacy-friendly image dataset purchasing via crowdsourcing[C]// Proc. of INFOCOM. ACM, 2018:2735-2743.
- [13]Gennaro R, Gentry C, Parno B. Non-interactive verifiable computing: Outsourcing computation to untrusted workers[C]//Proc. of CRYPTO. 2010: 465-482.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

量子计算五人谈

孙贤和
伊利诺理工大学

关键词：量子计算 量子通信

编者按：量子计算是一个大家都很感兴趣的课题，也是一个在快速发展中的课题。2019元旦假期期间，美国的一些学者在线上线下进行了关于量子计算的热烈讨论。我们邀请了孙贤和教授把他们的相关讨论进行了整理，在这里发表，以飨读者。

S : Sun Y : Yu W : Wang Q : Qiang X : Xu

近期的量子计算方向

S : @Y, 最近到华盛顿开会，听说你们申请到了一个量子计算的大项目，很为你们高兴。

Y : @S, 是的。我们最近刚申请到美国政府的一个做量子计算的大项目。美国政府最近推出了一个开展量子计算和量子通信的国家计划^[1]，我们这个项目是其中的一部分。

S : @Y, 国内关于量子计算和量子通信的报道和讨论特别多，有些离现实比较远。这是我2017年夏天访问美国洛斯·阿拉莫斯国家实验室(Los Alamos National Laboratory)之后写的一个关于量子计算机的解释(见附件“最简量子计算介绍”)。请你看一看，提提意见，也请谈一谈最新的进展。

S : 请注意，我在这个量子计算的简介中提到要实现Shor算法(Shor's algorithm)^[2]来分解一个大整数，如2048比特的整数，粗略估算需要4000个逻辑量子比特和2亿个物理量子比特，这显然是短期内无法实现的工程目标。那近期量子计算的研究目标是什么呢？

Y : 孙老师的“最简量子计算介绍”内容准确。但是从那时起量子计算又有了一些新的发展。为方便起见，我先简单介绍一下Shor算法。Shor算法是1994年麻省理工学院计算机学家彼得·肖尔(Peter Shor)^[1]提出的一种用量子计算机进行整数质因数分解的方法。它与已知的经典算法相比有指数级别的加速，而质因数分解本身也是广泛使用的RSA密码协议的基础，所以Shor算法可以对现有的公钥密码造成威胁。Shor算法主要利用了量子叠加态的“并行性”去分解质因数。基于量子计算机制造的现实情况，我们承认实现Shor算法确实短期内难以达到。同时我们也知道抗Shor算法攻击的基于Learning With Errors(LWE)^[3]问题的后量子密码体系正在建立。通过Shor算法以达到“量子霸权”(quantum computational supremacy)不是现阶段量子计算研究的主要目标。近期量子计算研究的一个主要目标是采用基于抽样的方法首次在50~100个左右物理比特的“量子计算机”上通过实验来展示经典计算机不能够完成的计算任务。也就是说，我们要在近期可以实现的量子计算设备上演示量子霸权。第二个目标是实

^[1] 肖尔时任贝尔实验室研究员，2003年加入MIT。

现最近提出的一些量子-经典混合策略的实用算法，用以解决一些实际问题，比如量子化学模拟，解决机器学习问题，达到一定程度的商用价值。现在面临的主要技术难点是量子比特的相干时间过短，以及量子电路里的噪声。加州理工学院的物理学家约翰·普瑞斯基尔 (John Preskill) 教授把现今处于的技术阶段称为含噪中尺度量子系统 (Noisy Intermediate-Scale Quantum, NISQ) 阶段^[4]。

S : @Y, 你说的第二个目标是可行的，也就是我们说的，以一种加速器的形式出现。

Y : 是的，现在考虑的计算模式就是让量子芯片像 GPU 一样，作为加速器和协处理器来使用。

如何在现有的量子计算机上实现量子霸权

S : 一台通用量子计算机需要 4000 个逻辑量子比特，但目前的量子计算只有 50~100 个物理比特，所以你们的小量子计算机只能解决很小一部分经典计算机可以解决的问题。因此它也只能以加速器的形式出现，帮助经典计算机解决问题。但在解决一些问题上，比经典计算机快并不能够证明经典计算机不能解决这些问题，这里面还有一个需要证明经典计算机无法解决这些问题的理论问题。你们的第一个目标——量子霸权的实验，具体是如何展现量子芯片速度的霸权优势的？

Y : 量子计算机在多项式时间内可以解决的问题类称为**有界错误量子多项式时间复杂性类** (Bounded-error Quantum Polynomial time, BQP)。而在证明 P 不等于 NP 之前，几乎没办法直接证明 BQP 严格真包含 P (或者 BPP²)。所以必须依赖一些复杂性的假设，比如之前计算机科学家设计的 RSA 加密就是依赖整数质因数分解十分困难的假设。但是 Shor 算法可以在多项式时间内解决整数质因数分解问题，所以整数质因数分解问题之前一直是量子计算超越经

典计算所依赖的例子。但因为 Shor 算法（分解 2048 比特的整数）需要超过 4000 个逻辑量子比特去解决，所以近期大家没有继续试图使用 Shor 算法来展示量子霸权，而是试图用 50~100 个物理比特的机器产生一个非常大的概率分布³，然后证明这个概率分布在经典计算机上因为算力不够而无法产生。这个方案有两个好处，第一是其依赖的复杂性假设非常强，假设的是**多项式时间层次** (Polynomial Hierarchy, PH) 从第三层开始不会坍缩。PH 是计算机科学家认为极为不可能互相等价 (即塌缩) 的一个复杂性类层次，所以这个假设比整数分解的困难性假设更强。第二是这个方案对错误的容忍度比较强，不需要可容错的量子机器来实现，适合现阶段的 NISQ 系统。这里需要说明的是，(量子采样) 这个计算任务本身是没有任何实用价值的，唯一的目的是为量子计算机量身定制一类问题，展示其计算能力。

S : 你这里还是有一个理论问题，因为需要证明这个分布无法用经典计算机产生。你们找到这样的例子了吗？

Y : 理论上已经证明了这样的分布在量子世界里是无处不在的。50~100 量子比特的量子态在哈尔测度 (Haar measure) 下必然会给经典计算机模拟不了的概率分布 (Porter-Thomas 分布)，所以现在实现量子霸权的方案之一是想用实验展示运行一段随机量子电路^[5]，在实验噪声控制得当的情况下，产生一个有 $2^{50} \sim 2^{100}$ 种不同输出结果的概率分布。而现今所有经典超级计算机都不可能从对输入的随机电路的描述中计算出整个输出的分布信息。

X : 最近已经有一项结果证明了 BQP 真包含 P，为什么不能直接应用在量子霸权的实验上？

Y : 我想你说的是然·拉茨 (Ran Raz) 等人的工作^[6]。虽然这项工作结论非常优美，但是不能帮助在实验上展示量子的加速，因为这个证明的方法是利用一个假想的运算能力强大的神谕 (oracle)，利用这个神谕，BQP 和 P 这两个复杂性类可以相对区分

² 有界错误概率多项式时间复杂性类，Bounded-error Probabilistic Polynomial time。

³ 即量子波色采样。

开⁴，但这个神谕在现实中是不存在的。

Q：量子化学计算的目的就是求解电子态的薛定谔方程，而量子计算本身也遵从薛定谔方程，所以非常适合用来做量子化学模拟。然而模拟后，我们要对量子计算机得到的量子态进行一一测量，这样的测量是否和直接测量要模拟的量子系统（比如说氢原子）的复杂度一样大？是否只是相当于把测量对象从原来的系统替换成了人造的量子系统？

Y：这是一个很好的问题。量子计算机的优势就是可以精确地对量子态进行操作和运算，在量子计算机中把体系的全部信息读出来是不可能的，但是利用相位估计算法 (phase estimation)^[7] 或者近期提出的 VQE(Variational Quantum Eigensolver) 算法^[8] 可以读出很多有用的信息，比如基态的能量。而在原系统中我们是不能进行这样的操作的。

S：@Y，你的第一个问题虽然是一个缩小的问题，但是从理论到实践是非常大的一步。首先，理论上证明不行的东西，在实际当中可能是可行的。理论只是说在问题大到一定程度的情况下，现在的机器会变得非常慢，以至于达到无法解决的程度。如果真找到了一个那么大的问题，这个问题在现有的量子计算机上又如何解决？要知道现在的量子计算机讨论的都是计算如何快，并没有讨论数据如何送上去。在实际中，量子计算机如何解决大数据的存储和传输问题？

Y：基于采样的量子霸权实验，这个问题的计算量非常大，但是问题的描述却十分简单，只需要包含一些量子门的电路图。量子计算机非常适合描述简单的问题，包括 Shor 算法解决的整数质因数分解问题、量子化学模拟都属于这类问题。在新兴的量子机器学习应用中，大数据的储存和传输就是很大的问题。现在量子科学家们正在积极寻找方法来解决这一瓶颈。

量子叠加态和量子通信

W：近来有人质疑量子计算机以及最近的量子通信实验都未真正实现量子纠缠和量子叠加态，这些质疑是否合理？

Y：对于判断量子叠加态和经典概率的区别，是有一个检验标准的，那就是贝尔不等式检验。贝尔不等式提出了一个可操作、可证伪的测试，具体形式是： $|P(xz) - P(yz)| \leq 1 + P(xy)$ ，其中 $P(ab)$ 为 a 、 b 事件同时发生的概率。任何没有量子特性的物理系统或者计算系统，是不能产生违背贝尔不等式的联合概率分布的。近期的量子计算实验以及量子通信实验均产生了违背贝尔不等式的概率分布^[9,10]。

S：你说量子通信实验产生了违背贝尔不等式的概率分布。中国现在大幅报道的量子通信实际上不是在用量子做数据传输，而是用量子做密钥分发。这个量子分发也不是用于量子加密，而是利用量子特性使得如果发密钥时被截取或者监听，分发的双方立刻就会知道。这里有两个问题：(1) 量子密钥分发是不是也产生了违背贝尔不等式的概率分布？(2) 针对 Shor 算法的量子密码体系已经建立，也就是说，实际上量子计算已经放弃了在密码破解上的应用。如果是这样的话，量子密钥分发与（后）量子密码一点关系都没有了，只是在为经典密码学服务。那现在的经典密码密钥分发是如何进行的呢？量子密钥分发的意义是什么呢？

Y：你的理解是对的。量子密钥分发确实只是确保密钥的安全性。实际上最早的密钥分发协议 BB84^[11] (Charles Bennett, 1984) 诞生于 1984 年，比 Shor 算法还要早十年。量子密钥分发基于纠缠态，所以肯定也可以产生违背贝尔不等式的概率分布，但是它利用的更多的是量子态的另外一个特性——不可克隆原理。不可克隆原理指出量子比特不能像经典比特一样随意地复制。每当一个量子被复制，原来的量子比特里的量子态必定被销毁，这是由量子理论的数学基础——线性代数决定的。不可克隆原理给量子计算带来了很多限制，但是 IBM 的查尔斯·本奈特 (Charles Bennett) 和吉勒斯·布拉萨德

⁴ 拉茨和塔尔 (Tal) 证明了如下结论：存在一个神谕 O ，使得 $\text{BQP}^O \neq \text{P}^O$ ，这与 $\text{BQP} \neq \text{P}$ 是不同的。

(Gilles Brassard) 却敏锐地发现了这个特点，生成和保管密钥——每当有人窃听，密钥发送者和接收者就会察觉自己的密钥分发效率下降了。需要指出的一点是，量子密钥分发要求必须有一条经典通信通道，这条经典通道里的信息默认是公开的。

Y：我对现在的经典密码密钥分发的情况不了解。

X：@Y，你提到美国政府最近推出了一个开展量子计算和量子通信的国家计划。但里面没有提到量子密码密钥分发。这是为什么？

Y：是的，美国的计划里没有关于密码密钥分发的讨论。

量子霸权和量子计算机的结构

X：回到量子算法的讨论，最近有人提出了一些经典算法可以达到和量子算法同样的加速。这是否意味着量子计算和经典计算机相比已经没有加速了？

Y：最近埃文·唐 (Ewin Tang) 等人提出了新的经典算法来求解用户推荐问题^[12] 以及矩阵求逆问题^[13]。他们的算法指数量级地缩短了求解的时间⁵，从复杂性角度上达到了和某类量子机器学习算法一样的速度，证明了量子算法在这些问题上没有优势。但是还有很多的量子算法与已知的经典算法相比有指数量级的优势，比如著名的 Shor 算法。埃文·唐的算法需要基于对输入矩阵的随机查询访问，而这类算法已经被证明（如果应用到整数质因数分解问题时）和 Shor 算法相比是有指数量级的差别的，所以新出现的算法在整数质因数分解等问题上是无望超过量子算法的。另外，新算法在解矩阵求逆问题上对于所求解的矩阵需要低秩等假设。新的经典算法虽然达到了同样的速度，但是它只能求解低秩的矩阵，而量子算法，例如 HHL 算法^[14]，没有低秩这一要求。

S：在理论上微观粒子的叠加态是天然存在的一种状态。但在任何时刻的任何测量我们只能测到一种状态。那么叠加态是如何用到量子计算上的？

Q：是的，这些态不可能同时被测出来。当只

有一个量子比特的时候，每一次测量只能测出一个具体的态，无所谓计算。但是当比特与比特间能耦合之后，我们在第一个量子比特进行输入，最后一个量子比特进行输出，中间过程不进行测量（当然这是个很简单的理想化的理论模型），那么中间的量子比特就能同时以无数种状态的情况进行相互作用，从而实现量子计算。至于中间的态如何“同时存在于无数种状态中”，并不是关心的重点。

S：@Q，讲到这里，你最好还是把量子计算机的基本结构介绍一下。它是图灵机吗？它还是冯·诺伊曼的结构吗？

Y：量子计算机基本结构称为量子图灵机，和 NP 图灵机非常相似，区别在于它可以有负的概率以及最后只能根据概率随机取出一个结果写入纸带。这个模型不太直观，所以不经常被提起。姚期智教授在 1993 年证明了量子图灵机和简单的量子线路模型等价^[15]。现在基本可以认为量子计算机就是一个可逆逻辑电路加上一些特殊的量子门。和冯·诺伊曼结构相比，量子计算机是没有外部储存设备的，其 CPU 就是内存。可以认为量子计算机直接在内存里进行运算。

S：如果没有外存，现在的量子计算机的数据存储和输入输出是如何完成的呢？

Y：粗略地讲，可以把每个量子比特看作一个 2 个经典比特的内存空间。这和经典的内存的区别在于，多个内存组合不是简单地相加而是张量积。 n 个量子比特储存的是 2^n 个经典比特的信息。在量子比特上进行量子门操作（物理上就是打激光、微波等操作）就等价于直接在内存空间里进行一个 $2^n \times 2^n$ 的矩阵运算。每个量子计算过程都可以看作在内存空间里的初始值上做矩阵和内存向量的乘法，然后以一定的概率读出最后内存空间的部分信息。在这个计算模型中，经典计算机的储存仍然充当了外存的作用。经典计算机可以通过控制量子门电路的参数来进行不同的矩阵运算。比如 Shor 算法，其中要分解的大整数的信息就是通过打的量子门来传递。

⁵ 求解时间从 $\text{poly}(n)$ 下降到 $\text{poly}(\log n)$ 。

对于这类算法，要传入的信息本身就是多项式规模，所以 I/O 操作不会是运算的瓶颈。比如要分解 9999 这个四位数，我们只要把它的二进制数位输入进去，而不是从 0 开始数一遍。对于量子化学问题也同样如此，只要把量子化学的相互作用的参数、系统大小确定好，I/O 不会成为限制。而对于另外一些问题，比如量子机器学习，如何突破 I/O 瓶颈还是个未解决的问题。值得注意的是，量子计算中的 I/O 瓶颈可以比经典计算更严重，经典计算中的 I/O 速度只是比 CPU 运算速度慢常数倍，而量子计算可以达到指数倍。

S：那怎么解决呢，除了尽量减少 I/O 和在调度方面做一些重叠以外。

Y：慢常数倍的问题不需要解决，比如量子化学；慢指数倍的暂时还不知道如何解决：比如量子机器学习如何把用户数据以大矩阵的形式装进量子计算机。孙老师是 I/O 方面的专家，请问一般 I/O 方面的问题解决的思路是什么？

S：数据传输的速度是由慢的一方决定的。既然量子计算机的储存就是经典计算机的储存，那么量子计算机的 I/O 速度也取决于经典计算机的储存速度。对数据密集型的应用来说，这是一个非常要命的问题。现在经典计算机的内存比外存快 400~4000 倍。解决的方法有两点：局部性和并行性。并行性包括重叠技术，当然，我们有很多不同的方法用以增加数据的局部性和并行性。例如，利用数据的特性去设计数据结构以增强局部性。量子计算的数据有什么特性，我不知道。但是在结构上，量子计算机的内存没有分级存储器体系 (memory hierarchy)，而分级存储是经典计算机的重要组成部分。

W：@S，听说你的研究小组今年（2018 年）也申请到了一个大项目，是做 I/O 系统的研究。你能介绍一下吗？

S：是的，我们最近有一个美国科学基金委 (National Science Foundation, NSF) 的项目，是做 I/O 系统的开发。基本的想法是，从高性能计算的 I/O 软件栈出发，打通上面的内存层和下面的储存层，建造一个 deep memory-storage hierarchy。

W：这个想法听起来很新颖。

S：我们这套系统有自己的理论和系统体系，但我们还是在经典计算机的框架下做的，与量子计算机没有交集。

总结

S：让我简单总结一下以上关于量子计算机的讨论：业界已经承认 4000 个逻辑量子比特（需要 2 亿物理量子比特支持）的通用量子计算机在短期内无法实现，现在的重点是实现 / 应用 50~100 个物理量子比特的量子计算机。也就是说，量子计算机只能解决一部分经典计算机能解决的问题。其应用与经典计算机结合，以加速器的形式出现。这样的应用在 5~10 年可以达到。因为这样的结合，并没有说量子计算机要做多少部分，也没有说量子计算机要做到多小。但另一方面，量子计算机的研究者“野心不死”，还要证明 50~100 个量子比特的量子计算机能达到量子霸权，也可以解决一些经典计算机解决不了的问题。理论上证明了这样的问题是存在的（虽然是一个人造的、不实际的问题）。但我认为，在现有的量子计算机上验证这个不实际的例子也不是一件容易的事，这包括现有量子计算机的硬件条件、软件条件和可能存在的 I/O 瓶颈。

Y：有一个小小的纠正，我们找的量子霸权的例子中没有 I/O 问题。

W：@Y，是的，你们避开了 I/O 问题。但你们要在真实的机器上实现目标还是很困难的。从理论到实践是一大步，现在最大的 IBM 量子计算机只有 50 个物理比特，据说还不能同时使用。你们的困难很多呀。

Y：量子计算是有很多困难。但是摩尔定律已经走到尽头了，经典计算的加速也变得越来越困难了。

S：“摩尔定律已经走到尽头了”这句话已经说了二十多年了，但我并不认为摩尔定律在可见的十年之内会走到头。我在今年（2018 年）10 月举办的中国计算机大会 (CNCC2018) 的“CPU 及 xPU 的未来之路”论坛上陈述了这一观点和论据。在那之后，

英特尔公布了他们的摩尔定律。在可见的将来，英特尔依然会依照摩尔定律来发展芯片。

Y：从理论上来讲，摩尔定律总有一天会结束。从理论上讲，量子计算机代表了未来。

S：从理论上讲，理论和实践是一回事；但从实践上讲，理论和实践并不是一回事。理论和实践中间隔着一大步，有时是无法逾越的。但我还是希望你们成功。你们成功解决实际问题的那一天，也就是 I/O 变成量子计算瓶颈的那一天，也许那时我们就可以合作了。

Y：在可见的将来，量子计算是以加速器的形式出现的。经典计算的成功，也是量子计算的成功。我也真心希望你们的项目成功。

S：祝你们成功，祝我们成功，祝大家成功。新年快乐！

Q, S, W, X, Y：新年快乐！ ■



孙贤和

2018CCF 海外杰出贡献奖获得者，2017 年 IEEE CS Golden Core Award 获得者，IEEE CS 旗舰期刊 *Transactions on Parallel and Distributed Systems* 副主编。伊利诺理工大学杰出教授。IEEE Fellow，美国阿贡国家实验室客座教授和伊利诺理工大学可扩展计算软件实验室主任。主要研究方向为高性能计算系统等。sun@iit.edu

参考文献

- [1] Lamar S. National Quantum Initiative Act[OL]. <https://www.congress.gov/bill/115th-congress/house-bill/622>.
- [2] Shor P. Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer[J]. *SIAM J.Sci.Statist.Comput.*,1994, 26(5):1484-1509.
- [3] Regev O. On lattices, learning with errors, random linear codes, and cryptography[C]// *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*. ACM Press, 2005: 84-93.
- [4] Preskill J. Quantum computing in the era of NISQ[OL]. (2018). <https://arxiv.org/abs/1801.00862>.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

附：

最简量子计算介绍

孙贤和

量子计算机是一个热门话题，经常被提起。在被人问了很多遍之后，我干脆把这个最简单的解释写下来与大家分享。

量子计算机有两个非常不一样的模型。一个是状态模型，一个是量子退火模型。量子计算基于量子比特 (qubits)，理论证明通用量子计算需要至少 4000 个逻辑量子比特。根据第二种模型做的量子计算机现在已经实现，已有厂家（如 D-Wave）在卖了。最新型的已经有 2000 个工作比特，已非常接近实用。但第二种模型只能做一种运算——模拟退火，并且不能证明比电子计算机快。用第一种模型做的量子计算机现在还都只有两位数量级的物理量子比特。要命的是，理论上已经证明如果有 4000 个逻辑比特，按照第一种模型造的量子计算机需要有 2 亿个物理量子比特，那是一个无法达到的高度。

所以如果有人告诉你量子计算机会比电子计算机快 1000 倍，并且量子计算机很快就可以实用，你不能说他在骗你，但是他肯定是在忽悠你。最好的可能是在不久的将来，第一种量子计算可以作为电子计算机的加速器出现，在实际运算中得到应用。

除了制造的问题，量子计算现在的理论证明都没有把 I/O 算在里面。如果把 I/O 的开销算在里面，量子计算机并不比电子计算机快。I/O 的问题是量子计算最重要的问题，比量子计算的算法还要重要。

量子计算投入小，曝光度高，所以大家都要做。但是大家还应当清楚地认识到量子计算离实际应用还有段距离。要真正取代电子计算机，只能说是有可能，但现在还看不到这种可能性。

区块链的极限

关键词：区块链 时间戳 哈希树 比特币

万 赞
休斯敦大学
特邀专栏作家

2018年12月11日，比特币从一年前的每枚2万美元降为3300美元，结束了几个世纪以来最大的金融泡沫。2017年一路疯涨的比特币泡沫受到全世界的关注，支撑它的区块链技术也随之声名鹊起，不少科技大咖将其和万维网技术相提并论，使其在短时间内成为投资领域的新宠。2018年随着比特币泡沫的破裂，区块链前景如何？是否仍有可能像有些媒体宣传的那样会取代互联网？还是最终会随着比特币泡沫的破裂而消失？

时间戳

比特币所依托的区块链技术在20年前就已出现。1991年商业互联网未出现时，民用互联网仅仅存在于研究机构、大学和少数科技公司之中，但互联网上的各种信息交换活动却相当活跃，尤其是电子文档的传阅。

1991年*Journal of Cryptology*上发表了一篇题为“如何给数字文档盖时间戳(How to Time-Stamp a Digital Document)”的论文^[1]。这篇论文探讨如何通过去中心化的方式给电子文档加盖有公信度的时间戳，使得任何人都可以准确分辨出一个电子文档的生成和修改时间。显然这一技术适用于很多法律方面的应用，比如为发明和专利的纠纷提供谁先提出

想法的依据。

该论文的两位作者哈伯(Stuart Haber)和斯托内塔(Scott Stornetta)都是科班出身。20世纪70年代哈伯毕业于哈佛大学数学专业，随后游学巴黎高等师范学院和斯坦福大学，最终在哥伦比亚大学读完密码学博士学位，成为这一领域的专家。斯托内塔比哈伯小10岁左右，20世纪80年代就读于斯坦福大学，直到获得物理学博士。两人一起写这篇论文的机缘是他们毕业后都去到当时热门的高科技公司Bellcore做研发工作^[1]。显然两人深厚的数理背景为区块链的发明提供了基础，成就了这篇区块链技术的开创性论文。

在这篇论文里，他们想通过盖时间戳的方式来准确区分电子文档的时间先后顺序。他们意识到单单通过一个具有权威性的时间戳服务器给文档加盖时间戳是不够的，因为服务器有很多潜在风险，比如被黑客攻击，发生计时故障，或者服务器的时间被拥有服务器控制权的人所更改等等，这些都可能导致时间戳的不准确，从而无法区分文档提交的顺序。他们认识到要想真正做到区分先后顺序，更可靠的方法是服务器把电子文档到达的先后顺序信息以不可更改的方式嵌入到每一个文档中。这样当用户收到盖上时间戳的文档后就多了一层验证机制。

针对这一想法，他们提出了两个解决方案，这两个方案的核心机制都是让后来提交的文档包含前

¹ 1983成立的Bellcore是贝尔电话公司被美国政府分拆后，众多的贝尔地区性公司共同投资成立的类似于贝尔实验室的研发机构。

面已经提交的文档的信息。这样一来，所有电子文档会根据生成的先后顺序逐渐形成一条文档链，后面的文档可以通过追溯的方式验证前面所有文档的时间顺序。即便有人想在时间戳上做手脚，也必须修改前面产生的文档的所有时间戳才行。所以对用户来说，要防止恶意篡改行为，只要保证修改所有文档的成本足够大，以至于几乎不可能实现就可以了，这就是区块链“链接”部分思想的缘起。而实现上述想法需要数字签名和哈希算法技术。

基于公钥加密技术的数字签名在 20 世纪 70 年代被提出，经过不断优化，到了 90 年代从实施效率来看已经非常完善。配备了公钥加密系统的时间戳服务器可以通过数字签名的方式给用户提交的文档加盖时间戳，在加盖时间戳之前的用户文档后面加上前面用户的文档信息就实现了文档链接。但是这里面还有一个效率问题。因为不同大小的电子文档会导致文件的上载和回传时间不同。如果前面用户的文档太长，一旦添加到后面用户的文档中，会导致后面用户文档存储空间的不必要的浪费等问题。于是哈希算法被引入。

哈希算法的特点是可以把不同长度的电子文档映射成标准长度的数文摘要，而且稍微变动文档的任何部分，都会导致哈希摘要的显著变化，尽管不能完全杜绝两个不同电子文档的摘要碰巧完全一样的情况，但产生这种冲突的概率还是很小的。每一个文档的哈希摘要都是固定长度的，服务器就可以把前面文档哈希摘要的已签名信息嵌入到后面文档中，然后给后面文档加盖时间戳并签名，再返回给用户。于是用户通过这一方式把自己的文档添加到了该服务器所创建的文档链中，可以通过后者回溯验证或者向第三方证明时间戳的准确性。

哈希树

哈伯和斯托内塔的论文虽然提出了文件链的思想和实现方式，但是在回溯验证方面仍然存在效率

问题。这是因为当用户对某一个文档的时间戳有怀疑时，他可以向前回溯相关文件，验证时间戳的可信性，但是回溯的效率会随着需要回溯文档链的增长而逐步下降。与此同时，当大量的平庸交易 (banal transaction) 也希望获得时间戳时，这种对每个交易进行单独链接方式的计算成本就变得非常高。于是他们与哥伦比亚大学的数学教授戴夫·拜耳 (Dave Bayer)² 合作，在 1993 年通过哈希树方式对文档链接技术的效率进行了完善^[2]，解决了批量处理电子文档或者电子交易的时间戳问题。

哈希树（又称 Merkle 树）是美国计算机专家默克 (Ralph Merkle) 在 1979 年提出并获得专利的一种计算机数据结构。它的基本概念是以二叉树的形式把需要加密的电子文档的哈希摘要存放到树的叶节点，然后将叶节点以上的每一层节点均以其子节点的哈希摘要进行重复哈希，直到根部。这样形成的一棵哈希树，无论有多少叶节点（对应电子文档的数量），只要任何一个节点被改动，那么哈希树的根节点也会发生变化，所以要验证该哈希树所包含的任何电子文档是否被篡改，我们只要验证包含该文档的哈希树的根节点，这就提高了验证效率。

利用哈希树的这一特点，他们把原来单个文档的直接链接方式转变为由一组文档形成哈希树的链接方式。这里的哈希树就是所谓的“区块”，并且在这种新的链接方式里，每个新文档需要记录的不再是前面文档的哈希摘要，而是所有与其相关，直到根节点的它所在哈希树的所有侧枝节点的哈希摘要。

更具有创新意义的是他们还提出了一种竞争机制，就是鼓励每个用户尽快计算出他们文档所在哈希树的根节点，最先算出根节点的用户通过向全网通告根节点来形成公共历史记录。后来的比特币显然受到这一思路的影响，用竞争计算方式（工作量证明）来产生新比特币区块，并且用奖励比特币的方式来鼓励竞争的比特币生成机制。后者则衍生出了一批专门从事创建区块的“挖矿”公司和与此相关的软硬件技术。

² 数学家拜耳的主要研究领域是纯数学，尤其是代数。他曾经受邀为描述数学家纳什的电影《美丽心灵》做学术指导。

比特币

1994年哈伯和斯托内塔离开 Bellcore 开始创业，成立了 Surety 公司，继续推广区块链技术。但是他们的推广并没有引起产业界和主流媒体的关注。到了 2008 年，神秘人物中本聪 (Satoshi Nakamoto) 声称开发了一款真正意义上的分布式加密虚拟货币。他为该虚拟币注册了网站 (bitcoin.org)，上载了虚拟币钱包和挖币软件，并用该软件在 2009 年 1 月 3 日挖掘创立了第一个比特币区块。于是比特币悄悄出现在互联网上。

虚拟货币作为网络支付的一种手段在互联网电子商务出现之前就已经被关注。推动这一领域发展的最初动力是小额支付问题。信用卡有最低交易成本，所以用信用卡来在线购买价值几美分的商品，交易成本就变得非常高，商家或者信用卡公司都倾向于拒绝这类交易。理想的解决方案显然是使用一种交易成本接近零的虚拟货币。针对这一需求，90 年代末互联网上曾经出现过多种虚拟货币尝试，并由此产生了研究和讨论虚拟货币的网上社群，俗称“币圈”。

但是虚拟货币的流通存在一个棘手的重复支付问题，也称“双花 (double spending)”问题。为了避免虚拟币持有人用同一币值进行多次支付，任何虚拟币系统都需要花费成本跟踪和标记每一个用户持有的币值。对跟踪成本的要求与接近零成本的交易显然有一定的内在矛盾。除了技术层面，虚拟货币还有一个设计层面的挑战。早期电子商务的创业者几乎全部崇尚哈耶克式的自由主义³，认为虚拟货

币的发行不应受包括政府在内的任何组织或者个人的控制，而应该像黄金一样具有稀缺性，需要通过一定的成本才能被生产出来，这样才能从根本上杜绝滥发贬值问题。这一成本需求恰好与前面提到的跟踪和标记功能在成本需求上有着一致性。所以虚拟货币的推崇者一直期盼能够有一种技术，以去中心化的方式花费一定成本跟踪和标记虚拟货币的使用，既解决了双花问题，又提供了对稀缺性的要求。

中本聪发现区块链恰好能够同时满足这几种需求。简单来看，区块链的时间戳可以准确地标记交易时间和先后顺序，避免同一币值被多次重复使用；用哈希树打包解决平庸交易的优化方案，可以直接应用到打包比特币交易为比特币区块；而用户通过竞争方式打包比特币交易进入新区块所产生的成本，也就是创建新区块的“挖矿”过程⁴，为比特币提供了稀缺性的特征。于是比特币借助区块链技术成为最能满足虚拟货币支持者的币种。

比特币在创立之初并没有引起币圈外媒体和投资界的关注⁵，这一事实从早期币值就可以看出。由于没有比较稳定的早期比特币和传统货币的兑换率，我们只能从一些有限的交易中推算出当时每枚比特币的币值不到 1 美分⁶。尽管如此，由于比特币所使用的区块链技术使它较为成功地解决了前面提到的虚拟货币在发行和流通层面的两大挑战，逐渐吸引了越来越多的用户和炒家。最终通过马太效应从众多虚拟币中脱颖而出，并在 2017 年由于炒家的过渡投机，造成金融史上出现超过“荷兰郁金香泡沫”⁷的最大金融泡沫。不过区块链技术却借

³ 哈耶克式的自由主义指的是一个人不受制于另一个人或另一些人因专断意志而产生强制的状态，重视个人自由和自由的价值，反对国家对个人的强制，对当今世界各国政治及法治建设起到了十分重要的借鉴作用。

⁴ “挖矿”又称“工作量证明”，是通过让参与区块链的服务器不断猜测符合规定条件的新区块头的固定字段来实现。具体来讲，就是不停地变更区块头中的随机数，并对每次变更后的区块头做双重 SHA256 运算，然后将结果值与当前网络的目标值做对比，如果小于目标值，则解题成功，工作量证明完成。

⁵ 中本聪在挖掘了一百万枚比特币后销声匿迹，把比特币网站和管理任务留给了比特币程序开发员 Gavin Andresen，后者同时接管了比特币协议的维护工作。

⁶ 最早的比特币购物记录发生在 2010 年 5 月 22 日，一个佛罗里达州的程序员在币圈论坛里表示愿意用 1 万比特币换取两块比萨饼。于是一个英国圈友花了 25 美元从网上订了两个比萨饼送到了他家里，获得了这 1 万枚比特币。按照这一价值估算，当时的一枚比特币大约可以兑换 0.0025 美元。

此契机引起世人的瞩目。

公信与应用

区块链技术最重要的特点是它能够给所有参与者提供一种公信力。比如比特币的使用者相信他们嵌入到区块链中的比特币币值只有他们可以使用，而且是可以得到准确验证的。是这种对比特币区块链的信任使得不断有人愿意购买比特币和接受比特币。

公信力的传统来源是权威机构，譬如政府或者是 RSA⁸这样的专职网络信用授权公司。这种公信形式的特点是每一个参与者将对公信对象的信任，委托或者转移给提供公信力的权威机构。譬如国民因为信任主权政府而信任政府所发行纸币的购买能力，网民因为信任 RSA 而信任该公司所发放的对各种网络公司和个人的身份认证。

传统公信形式的特点是效率高，成本低。无论是直接发行货币还是直接发布数字证书，都要比区块链的解决方案来的直接。但是这种公信形式的缺点也很明显，就是风险性大。这里的风险性包括系统风险和道德风险。系统风险是指在公信委托或者转移体系的设计中存在可以被不法分子有机可乘的漏洞，比如若黑客攻入 RSA 认证服务器获得根认证密钥，就有可能导致整个互联网电子商务系统的认证系统紊乱和崩溃，使得正常的电子交易无法进行。道德风险是指被委托的权威机构因为集中掌握了公信权而利用该权利，做出有利于自身但是损害参与者的行为所带来的风险，比如政府滥发货币引发通货膨胀，损害普通居民的经济利益。

尽管传统公信形式的风险时刻存在而且不容忽视，但在信息技术普及之前，人类社会并没有找到一个很好的替代形式。纸币是世界各国普遍采用的公信形式，唯一例外是雅浦岛的石币^[3]，但是其因为物理介质的局限性只能在很小的社会范围内使用。

信息技术尤其是互联网技术的普及为突破传统

公信形式提供了重要的契机。哈伯等学者发明时间戳和用哈希树打包交易的初衷是解决传统公信形式的系统风险，而中本聪则试图用比特币和区块链解决政府所发行的法定纸币所带来的道德风险，这也是哈耶克的初衷。

区块链所提供的新的公信形式可以应用到很多方面。如果把区块链的核心技术进行分解分析，可以进一步发现，比特币利用的主要时时间戳功能，并以此来杜绝“双花”企图。如果我们保留链接加密理念，同时提供参与者的地理位置信息，区块链就可以应用到流通领域的供应链管理中。比如受很多消费者崇尚的有机食品和农产品存在不同程度的造假现象，尽管推出这些产品的公司提供了所谓的认证，但越来越多的权威认证机构的信任缺失事件，使得越来越多的人对货架上商品的产地真实性持怀疑态度。针对这种现象，学术界和企业界提出了用包含地理位置信息的区块链来链接一件商品的整个供应链参与者的策略，使消费者可以通过使用智能设备扫描商品上二维码等手段，验证一件商品从产地到批发到零售所经过的所有地区和参与者信息。因为这一信息是每个经过身份验证的参与者自愿提供的，其可信度要比仅由商家或者权威组织提供的供应链信息要更加可信^[4]。

区块链公信的另一个重要的应用领域是法律合同。前面提到时间戳是为了解决网络文档的首发时间辨别。显然在法律领域除了专利申请外，还有很多分支需要时间先后顺序的辨别。其中最普遍并且需求量最大的是遗嘱建立和修改。西方国家的遗嘱管理通常通过律师进行，中国的遗嘱则是通过公证部门办理。区块链技术为遗嘱的创建和管理提供了新的途径，因为它不但可以通过公密和数字签名保障遗嘱文字的完整和准确，而且通过时间戳和链接顺序保障了遗嘱修改和重立后的最终版本的有效性。

除了替代传统法律合同的公信，区块链还可以在这一领域通过加强应用提供智能合约^[5]。建立在

⁷ 荷兰郁金香泡沫是人类史上第一次有记载的金融泡沫经济，此事间接导致了作为当时欧洲金融中心——荷兰的衰落。

⁸ 世界级信息安全解决方案的主要提供商，帮助世界领先企业成功解决最复杂敏感的安全问题。

区块链技术上的智能合约，可以在保护和甄别合同完整性和有效性的同时，提供一定的自动执行功能，当然前提是合约的执行条件和执行通道能够预先编写到区块链和镶嵌到系统中。

平台的限制

尽管区块链技术有着非常广泛的应用场景，但它是否有可能突破公信应用成为一个可以容纳其他应用的平台技术？答案是否定的。

信息领域里平台型技术的代表是互联网和万维网。互联网是各种不同传输协议的计算机网络之间联网的统称。在万维网出现之前，这些使用不同协议的网络通过路由和网关实现相互之间的通讯，从而为计算机上的各种应用提供一个可以交流的平台。万维网则是更进一步，将互联网各种网络协议的内容和应用通过万维网协议进行整合，方便了用户。当万维网成为主流后，其他类似功能的网络协议逐渐被废弃，而更多的应用开始搭建到万维网上，从而进一步确定了万维网的平台技术地位。

万维网之后，出现了多个与万维网互补或者是搭建于其上的围墙花园式平台，比如基于无线应用协议 (WAP) 的移动平台和以脸书与微信为代表的社交商用平台。这些后来出现的平台技术与万维网都有一个共同特点，平台本身是技术中性 (techno-neutral) 和数据结构盲视 (data structure blind) 的。比如无论是万维网还是移动网络，任何应用中的数据在这两个平台上的流动无须加密，但是企业可以根据需要在传输时将数据加密。同时平台可以完全根据应用的需要，接纳任何数据结构。脸书和微信虽然依托于万维网或者是移动网络之上，但它们依然能够提供一个具备以上特点的专门平台环境，使开发人员不但可以在其之上开发出各种应用，而且可以充分利用它们各自拥有的用户信息。

区块链则不具备上述的包容性。首先，区块链的参与者必须将所有入链的数据加密和签名才能保证真实性，不具技术中性要求；其次，区块链的数据组织方式是固定的链表 (Linked List) 数据结构，

这使得需要使用其他数据结构的应用程序无法利用区块链做为平台。所以至少从目前的区块链架构和数据组织方式来看，它无法成为一种普遍意义上的平台技术。

未来与局限

显然区块链在取代传统公信方面有现实意义。比如法律文档，合同，财务交易记录等等，我们可以通过区块链技术来降低风险和提高透明度。但是我们也需要认识到区块链技术的局限，尤其是目前的几个重要应用领域都存在难以克服的障碍。

智能合约是区块链的重要应用领域。比特币的 Script 和以太坊的 Solidity 都是可以提供智能合约的专门语言，而且以太坊的 Solidity 是图灵完备语言，可以用来编写复杂的合约。但是合约越复杂，执行条件就越难分析，计算机可以理解的方式就有可能出现偏差，所以到目前为止，我们仍然难以在区块链主导的智能合约领域有实质性突破^[6]。

目前已经降到 4000 美元以内的比特币存在能源消耗所带来的造币成本挑战。从 2015 年开始最近一轮的比特币升值，导致大批公司和个人将计算资源投入到创建区块的工作量证明竞争计算中。但是无论多少个竞争者参与，每一个新的区块只能有一个创造者，这就意味着每一个新区块面临潜在成本的不断增加和海量计算资源的浪费。根据美国知名华尔街调研机构 Fundstrat 的估计，目前新比特币的区块链计算成本已经达到 4500 美元，超过了比特币市场价格。针对这一问题，有专家提出用股权证明方式来替代工作量证明创建新区块。这种方式虽然减轻了能源消耗，却增加了“双花”风险，其应用前景有待观察。

除了应用层面的挑战，区块链还有一个根本的局限性。它所提供的公信的维护成本要高于传统公信的维护成本。我们从比特币的应用中可以看出，区块链不是一种能够完全去中心化的技术，而是一种分布中心化 (distributed centralization) 技术。虽然它不依赖于唯一中心服务器存在，却要通过多个服

务器的不断竞争与合作，来为参与到链里的成员提供相应的区块产生和维系服务，这种高成本合作关系的维系是区块链存在的必要条件，所以维护区块链所需要的最低经济成本门槛，限制了更多传统公信领域转化到区块链应用的可能性。 ■



万 赞

CCCF 特邀专栏作家。美国休斯敦大学维多利亚校区教授。主要研究方向为电子商务和互联网应用。著有《电商进化史》一书（机械工业出版社 2015 年出版）。
wany@uhv.edu

参考文献

- [1] Haber S, Stornetta W S. How to time-stamp a digital document[J]. *Journal of Cryptol.* 1991;3(2): 99-111.
- [2] Bayer D, Haber S, Stornetta W S. Improving the efficiency and reliability of digital time-stamping[J]. *Sequences II: Methods in Communication, Security, and Computer Science*, 1993:329-334.
- [3] Furness W H. *Yap of the Carolines*. JB Lippincott Company, 1910.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

CCF 表彰 2018 年度优秀专业委员会

在 2019 年 1 月 19 日举行的 2018 CCF 颁奖大会上，中国计算机学会 (CCF) 正副理事长向获得优秀专委奖的中文信息技术专委、人工智能与模式识别专委、高性能计算专委、软件工程 / 系统软件专委、数据库专委颁发了证书。CCF 理事长向获得专项奖的大数据专家委员会、多媒体技术专委、体系结构专委、抗恶劣环境计算机专委、计算机视觉专委、区块链专委颁发了证书。

中文信息技术专委管理运营规范，举办的学术会议国际化特点鲜明，组织了语义分析评测和竞赛等特色活动；**人工智能与模式识别专委**凝聚力强，组团参加 CNCC 并组织技术论坛，成功举办了 CCF-ICAI 国际会议，在人工智能的学术推广和服务上做了大量工作，提升了 CCF 在国际人工智能学术界的影响力；**高性能计算专委**的学术年会参会人数再创新高，吸引众多国内外厂商和科研院所参展，技术论坛反响热烈，积极组织“走进高校”等活动，推动我国 HPC 的发展；**软件工程 / 系统软件**两个专委成功联合举办 NASAC 2018、软件工程 50 周年等多个学术活动，积极促进学术产业交流；**数据库专委**组织出版的 DSE 国际期刊取得良好进展，长期举办 VLDB 暑期学校效果好，还举办了《数据库系统概论》出版 35 周年等多个特色活动。

大数据专家委员会通过举办大数据学术大会、大数据技术大会、大数据创新创业大赛等活动，发布 2019 年大数据发展趋势预测报告，出版大数据教材系列丛书，极大提升了专委在业界的影响力，获得学术产业合作奖；**多媒体技术专委**连续 6 年积极组织撰写 CCF 技术发展报告，进一步打造 China MM 学术年会品牌，与 IEEE、ACM 进行多方面学术交流，扩大国际影响力，获得国际合作奖；**体系结构专委**与企业合作组织论坛及挑战赛等特色活动，不断努力提升国际会议和学术年会在本领域的影响力，获得国际合作奖；**抗恶劣环境计算机专委**成功打造“自主可控计算机大会”品牌，广受业界关注和好评，走进企业进行深度调研，促进技术进步和企业合作，获得特色活动奖；**计算机视觉专委**由于学术会议影响力大且吸引众多知名企业赞助，积极开展“走进高校”和“走进企业”活动，在扩大专委和领域的影响力方面做了很多工作和努力，获得年度特别奖；**区块链专委**为 2018 年新组建专委，本可以不参加今年的专委评估工作，但该专委在成立不足一年的时间内成功举办了包括学术年会在内的多项学术活动，对总部的财务贡献突出，因此获得特别贡献奖。

机器阅读理解： 如何让计算机读懂文章

朱晨光
微软公司

关键词：阅读理解 上下文相关词向量 注意力机制

机器阅读理解任务

在2013年之前，自然语言处理(Natural Language Processing, NLP)研究中的主要任务集中在对词和句子的理解，例如词向量、句法分析、歧义消除等。而对于更复杂的结构，例如段落和文章，因其分析难度大而鲜有相关研究。近年来，深度学习在自然语言处理方面的突飞猛进，使针对句群和段落的语义分析成为可能。

基于人类的认知，判断阅读者是否理解一篇文章最直接的方式就是进行问答考核，即给定文章和与之相关的问题，评判阅读者给出的答案是否正确。因此，机器阅读理解多以问答形式来判断人工智能是否理解文章。机器阅读理解在工业界有着广泛的应用，例如：搜索引擎可以根据用户输入的查询来找到相关文档并精确给出答案；客服对话机器人可以根据用户的问题找到解决问题的文档并显示出具体的解决步骤等。

基于机器阅读理解的重要应用价值，在2013年微软发布McTest数据集^[1]之后，有十多个大规模机器阅读理解任务的数据集产生。

根据给定语料的范围，可以将机器阅读理解的任务分为两大类：单段落问答任务和多段落问答任

务。单段落问答任务是给定一个段落，其长度通常在数十到数百词之间，对于一个和段落相关的问题，算法需要在段落中找到对应的答案。图1是“讯飞杯”中文机器阅读理解评测中的一个样例。这类的数据集有McTest^[1]、SQuAD^[2]、CoQA^[3]、RACE^[4]等。多段落回答任务通常给定一个大的语料库，包含许多

段落

工商协进会报告，12月消费者信心指数上升到78.1，明显高于11月的72。另据《华尔街日报》报道，2013年是1995年以来美国股市表现最好的一年。这一年里，投资美国股市的明智做法是追着“傻钱”跑。所谓的“傻钱”策略，其实就是买入并持有美国股票这样的普通组合。这个策略要比对冲基金和其他专业投资者使用的更为复杂的投资方法效果好得多。

问题1：什么是傻钱策略？

答案：买入并持有美国股票这样的普通组合。

问题2：12月的消费者信心指数是多少？

答案：78.1。

问题3：消费者信心指数由什么机构发布？

答案：工商协进会。

图1 机器学习阅读理解样例¹

¹ 来源：<http://world.people.com.cn/n/2014/0101/c1002-23995935.html>。

文章与段落。对于一个问题，算法需要利用检索定位到答案可能存在的段落，再进行回答，这使得对于模型准确度的要求大大增加。相关数据集有 MS MARCO^[5]、ARC^[6]等。

根据答案的形式，机器阅读理解任务可分为段落中连续片段和多项选择两种。对于段落中连续片段任务，答案一定是段落中一段连续的文本，即模型只需要给出答案在给定段落中的起止位置，如图1。代表数据集有 SQuAD^[2]、CoQA^[3]。对于此类阅读理解任务，评判的标准一般为正确答案与模型给出答案之间分词后的精确率、召回率和 F1 分数（如图 2）。多项选择任务即给定若干备选项，算法需要选出一个或多个正确选项。代表数据集有 McTest^[1]、ARC^[6]等。

当前机器阅读理解任务的答案形式大多数为以

正确答案：买入 并 持有 美国 股票 这样 的 普通 组合
模型答案：其实 就是 买入 并 持有 美国 股票
精确率：5 / 7 = 0.71
召回率：5 / 9 = 0.56
F1：0.63

图 2 段落中连续片段类型的阅读理解任务指标计算

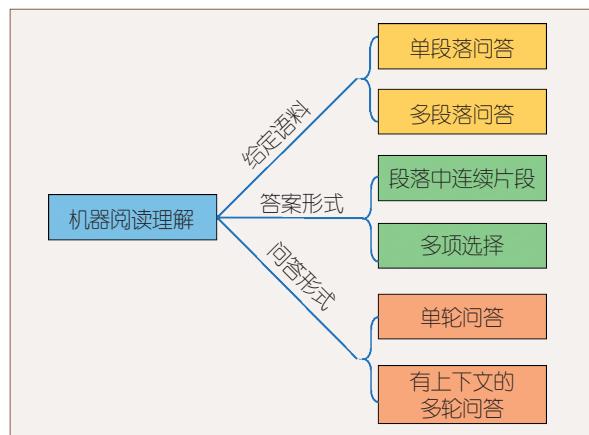


图 3 机器阅读理解任务的分类

上两种，形式均为限定的。其原因在于，判断自由形式的答案与正确答案是否语义一致本身是很困难的问题。

根据问答形式分类，可以将阅读理解任务分为单轮问答和有上下文的多轮问答两类。单轮问答中，不同轮的问题和答案之间没有相关性，可以独立求解。大部分阅读理解任务属于该类型。而从 2018 年开始，有上下文的多轮问答任务逐渐引起大家的关注。这类任务中，邻近轮的问题和答案之间存在相关性，即回答第 $N+1$ 轮的问题有可能需要依据第 N 轮及之前的问题和答案。这种形式的阅读理解任务更符合人与人之间对话的过程。相关数据集有 CoQA^[3]、QuAC^[7] 等。

模型求解

随着深度学习的不断发展，现在的机器阅读理解算法大多数是端到端深度学习模型：输入为原始的段落和问题文本，中间是可以求导优化的网络模型，而输出即为答案。这种形式给模型的建立和优化带来了极大的便利。

机器阅读理解模型的核心是建立给定段落和问题之间的语义联系，除了需要分别对段落和问题进行语义分析，也需要在段落中寻找与问题相关的片段。因此机器阅读理解模型至少需要三个核心模块：(1) 对段落 / 问题进行有上下文的语义分析；(2) 在段落的不同片段和问题之间计算相关度，并更新语义分析结果；(3) 根据相关度生成答案。

上下文语义分析

对于输入的原始文本，模型需要先经过分词，然后将每个词转化成固定长度的词向量进行分析。自然语言处理中比较通用的词向量表示有 GloVe^[8]、word2Vec^[9] 等。例如 GloVe 可以将任意一个词转化成为 300 维² 的实数向量。

由于词向量表示是固定的，而一个词在不同的

² 模型通过训练将词映射成 k 维实数向量， k 一般为模型中的超参数。

语境下可能有不一样的语义，因此必须根据其上下文计算出语境相关的语义表示。通用的方法是在词向量上采用循环神经网络(RNN)，通过信息的流动获得每个词在上下文中的向量表示。

自2018年以来，用于计算上下文相关词向量的预训练语言模型，因其优秀的性能而引起了研究人员的广泛关注。这类语言模型通过在大规模语料库上的训练来获得网络结构的参数。这样的网络结构产生的上下文相关词向量，在许多自然语言处理任务中取得了非常不错的结果。这类预训练语言模型包括ELMo^[10]和BERT模型^[11]。BERT模型基于Google提出的Transformer编码器，通过在海量语料库中的无监督学习，获得的预训练上下文相关词向量编码网络在11种不同的自然语言处理测试中获得最佳成绩，尤其是在机器阅读理解任务SQuAD v1.0中，首次在所有指标上超越了人类。因此，最新的机器阅读理解模型在上下文语义分析模块中基

本都使用了ELMo或BERT，然后使用循环神经网络获得词向量。

计算段落片段与问题相关度

由于问题与段落相关，对于段落的语义分析一定要基于对问题的理解。最基础的方法是，对于段落中的每个词，用一个比特位来表示它是否在问题中出现。而在深度学习中，使用注意力机制可以得到更好的效果。

给定文本A和文本B，注意力机制可以计算出基于对B的语义分析，A中每个词语的词向量表示。算法1给出了一种基于内积的注意力机制计算方法。在注意力机制中，文本A每个词语的词向量表示是文本B词向量的一种线性组合，该组合的系数来自A和B每个词对之间的匹配程度。

除此之外，自注意力机制可以生成一个文本基于自身词对之间匹配程度的词向量表示，即A=B。

算法1 基于内积的注意力机制

输入：

文本A的m个单词的向量表示： $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$

文本B的n个单词的向量表示： $(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$

步骤1：计算A中第i个词和B中第j个词的匹配程度 $s_{ij} = \mathbf{x}_i^T \mathbf{y}_j$

步骤2：利用softmax归一化得到注意力相关系数 $\alpha_{ij} = \frac{\exp(s_{ij})}{\sum_k \exp(s_{ik})}$

输出：文本A中第i个单词基于对B的语义分析得到的向量表示： $\mathbf{z}_i = \sum_j \alpha_{ij} \mathbf{y}_j$

算法2 段落片段式答案的生成方法

输入：

段落的m个单词的向量表示： $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m)$

问题的n个单词的向量表示： $(\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$

步骤1：线性组合得出问题的单向量表示： $\mathbf{q} = \sum_j \alpha_j \mathbf{y}_j, \alpha_j \propto \exp(\mathbf{u}^T \mathbf{y}_j)$

步骤2：计算答案在段落每个词语位置开始的概率 $p_i^s \propto \exp(\mathbf{x}_i^T \mathbf{W}^s \mathbf{q}), 1 \leq i \leq m$

步骤3：计算答案在段落每个词语位置结束的概率 $p_i^e \propto \exp(\mathbf{x}_i^T \mathbf{W}^e \mathbf{q}), 1 \leq i \leq m$

输出：

选择概率最大的区间作为答案： $(i^*, j^*) = \underset{i \leq j < i+maxlen}{\operatorname{argmax}} p_i^s p_j^e$ ，maxlen为预设的答案最大长度

其意义在于，循环神经网络在长文本中会出现影响力递减的情况：长文本中相隔较远的词之间很难互通信息。而自注意力机制通过计算单词两两之间的匹配程度，打破距离的限制，从而获得更精确的文本理解。

答案生成

对于多项选择类型的答案，可以通过神经网络中的线性变换层生成每个选项的概率，并通过交叉熵求导优化。对于段落片段类型的答案，一般是计算段落中每个位置作为答案开始和结束的概率，并选择概率最大的区间输出。算法 2 给出了一个典型的段落片段类型答案的生成方法，其中 u, W^s, W^e 均为参数。

在实际模型中，以上三个核心模块均会有不同变种，例如更多的循环神经网络层数、注意力机制的反复使用等。图 4 展示了最近在机器阅读理解多轮问答数据集 CoQA 上获得第一名的 SDNet 模型^[12]。

该模型使用了 BERT 输出层的线性组合以及复合注意力层的模块，获得了很好效果。

值得一提的是，在影响力很大的机器阅读理解竞赛 SQuAD 中，微软亚洲研究院、国防科技大学、哈尔滨工业大学与讯飞联合实验室等中国团队排在前列³（见图 5）。

挑战

2018 年初，在斯坦福大学推出的机器阅读理解数据集 SQuAD 上，来自微软和阿里巴巴的研究团队均在精确匹配程度指标上超越了人类，相关媒体也发出了“人工智能已经在阅读理解任务上战胜了人类”的口号。但是，相关研究表明，现在的机器阅读理解模型很大程度上依赖于简单的文本匹配。一旦给段落加入干扰信号，模型表现会大幅下降。在文献 [13] 中，关于 SQuAD 数据集上表现最好的 16 个机器阅读理解模型的实验表明，在段落中加入

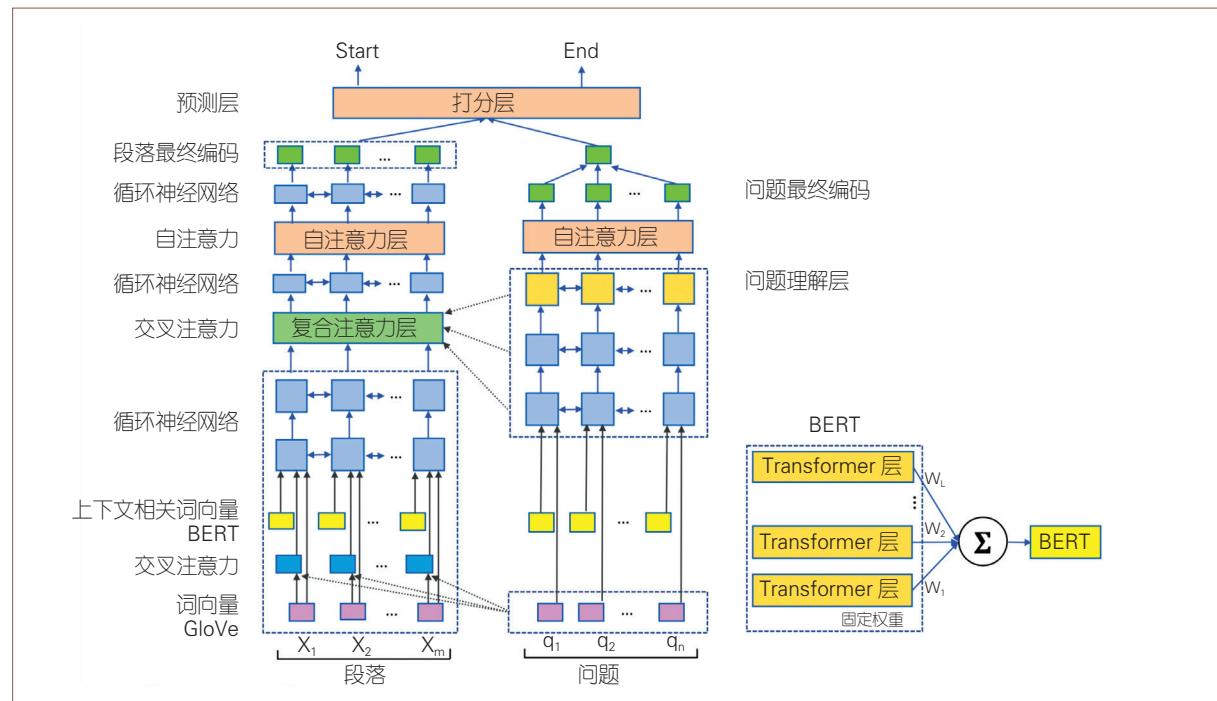


图 4 SDNet 模型

³ 来源：<https://rajpurkar.github.io/SQuAD-explorer/>。

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2 Oct 05, 2018	BERT (single model) Google AI Language https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
2 Sep 27, 2018	nlnet (ensemble) Microsoft Research Asia	85.954	91.677
3 Jul 11, 2018	QAQNet (ensemble) Google Brain & CMU	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) Microsoft Research Asia	84.003	90.147
5 Mar 19, 2018	QAQNet (ensemble) Google Brain & CMU	83.877	89.737
5 Sep 09, 2018	nlnet (single model) Microsoft Research Asia	83.468	90.133
5 Jun 20, 2018	MARS (ensemble) YUANFUDAO research NLP	83.982	89.796
6 Sep 01, 2018	MARS (single model) YUANFUDAO research NLP	83.185	89.547
7 Jan 03, 2018	r-net+ (ensemble) Microsoft Research Asia	82.650	88.493
7 May 09, 2018	MARS (single model) YUANFUDAO research NLP	82.587	88.880
7 Feb 19, 2018	Reinforced Mnemonic Reader + A2D (ensemble model) Microsoft Research Asia & NUDT	82.849	88.764
7 Jan 22, 2018	Hybrid AoA Reader (ensemble) Joint Laboratory of HIT and iFLYTEK Research	82.482	89.281

图 5 在 SQuAD 1.0 上排名前列的模型成绩与相关单位

一个干扰句（干扰句和正确答案所在句的文本匹配度高，但关键词不一样）会使得所有模型的 F1 指标均下降 20%~40%，而人类实验者的指标只下降 3%。该结果表明，现有的机器学习模型仍远没有达到人类理解文本的精确程度。如何加强模型的深度理解能力是这一领域的重要课题之一。

无监督学习 现有的阅读理解数据集规模与大规模语料库相比依然很小，如何使阅读理解模型不完全依赖于标注数据是一个很有意义的方向。

推理能力 当前的大多数数据集不需要多层推理便可获得正确答案，即事实型问答。人类特有的推理归纳能力还没有在机器阅读理解中得到体现。实现推理能力是文本理解中的重要课题。

文本表示 自然语言处理中关于词向量和句向量已有大量研究。而对于段落和长文本，一直以来没有很好的向量表示方法。由于机器阅读理解任务与段落相关，利用其问答数据建立有效的文本表示将有助于其他任务的处理。

最近几年，阅读理解作为应用广泛而难度较大的课题，吸引了许多研究者的关注。相应的数据集和竞赛层出不穷，带动了更新更通用的深度学习模块的研究（例如预训练语言模型 ELMo 和 BERT）。深度学习在这一领域取得了长足进步，但离人类水平还有很大差距，真正的智能文本理解依然是一个巨大的挑战。■



朱晨光

美国微软公司研究员，斯坦福大学计算机博士。主要研究方向为自然语言处理，包括机器阅读理解、人机对话及词嵌入。
zcg.cs60@gmail.com

参考文献

- [1] Richardson M, Burges C J, Renshaw E. McTest: A challenge dataset for the open-domain machine comprehension of text[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013: 193-203.
- [2] Rajpurkar P, Zhang J, Lopyrev K, et al. Squad: 100,000+ questions for machine comprehension of text[OL]. arXiv preprint arXiv:1606.05250.2016.
- [3] Reddy S, Chen D, Manning C D. CoQA: A conversational question answering challenge[OL]. arXiv preprint arXiv:1808.07042.2018.
- [4] Lai G, Xie Q, Liu H, et al. Race: Large-scale reading comprehension dataset from examinations[OL]. arXiv preprint arXiv:1704.04683.2017.
- [5] Nguyen T, Rosenberg M, Song X, et al. MS MARCO: A human generated machine reading comprehension dataset[OL]. arXiv preprint arXiv:1611.09268.2016.
- [6] Clark P, Cowhey I, Etzioni O, et al. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge[OL]. arXiv preprint arXiv:1803.05457.2018.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>

CCF TC

CCF 专委发展 & 交流会在京举行

2019年1月3~4日,CCF专业委员会发展研讨暨工作会议在CCF总部举行。CCF理事长高文、秘书长杜子德、专委工委主任胡事民、专委工委9位委员,以及CCF下属36个专委(专业组)60余位主任、副主任及正副秘书长出席了本次会议。会议由胡事民主持。

胡事民从专委基本概况、学术活动、企业合作、国际合作、服务能力建设,以及专委学术活动调研的反馈情况等几个方面对2018年度专委发展情况作了详尽的报告。

CCF软件工程专委主任金芝、中文信息技术专委秘书长赵东岩、区块链专委主任斯雪明,分享了各自专委在学术年会、企业合作与赞助、新专委组织活动等方面的经验。



杜子德秘书长指出,专委是学会开展学术活动的主体,在CCF占据重要地位,近年来不断进步。专委应与总部紧密结合,构建CCF架构下的统一品牌,专业化、规范化地管理和运营,对学会作出更大的贡献。

高文理事长指出,CCF的会员数量和质量都非常重要,希望专委能够吸引更多计算机领域专业人士参与CCF学术活动,加入CCF。他还号召专委在期刊出版和组织国际会议方面做出努力,在提高CCF学术影响力和国际影响力方面作出积极的贡献。

在专委工作交流环节,各专委负责人从2018年度专委目标的完成情况,专委的特色、亮点与经验,存在的不足和改进措施等几方面作了报告。

CCF教育工委走进高校组召开2019年工作会议

2019年1月11日,CCF教育工委走进高校工作组召开2019年第一次工作会议。CCF秘书长杜子德,CCF教育工委主任杜小勇、副主任高小鹏、主任助理张孝、委员孟祥旭等11人参加会议。

会议重申了CCF教育工委“走进高校”工作组的工作定位和工作目标,回顾了试点高校的情况。多数试点高校存在培养目标缺乏特色、培养方案的制订未能体现OBE(面向产出的教育)理念、毕业要求的分解和支撑缺乏合理性、课程体系建设有待完善、缺少持续改进机制等问题。2019年走进高校工作组将完成第二轮入校,汇总共性问题、组织交流研讨会,并征集《计算机科学与技术专业培养方案编制指南》第二版的修订意见,补充相关案例并完成修订工作。杜子德肯定了CCF教育工委“走进高校”工作组2018年的工作,并对2019年工作提出建议,希望加大宣传力度,扩大项目影响力,加强与国际相关组织间的沟通交流。



The CS David 专栏

CCCF 2019 年第 2 期

动物计算

关键词：动物 - 计算机交互 农业 智能

这本会议论文集附带了一条注解说，“你会对它感兴趣的”。其实不需要加这样的评论，因为我对几乎所有计算机领域，特别是那些似乎来自计算机边缘领域的报道都感兴趣。当我终于有机会打开它的时候，我还是惊讶地笑了出来，这本论文集来自一个动物 - 计算机交互 (Animal Computer Interaction, ACI) 会议——该会议似乎在模仿著名且参会人数众多的人机交互 (Human Computer Interaction, HCI) 会议。

我曾经参加过人机交互会议，我发现这是一个激动人心的会议，充满了活力和新想法。由于与会人数太多，包括我在内的许多报告人只能作限时 60 秒的发言。与会人员大约有一万名，挤爆了会场的 Wi-Fi 和本地手机网络。为了获得一个稳定的网络连接，我不得不去半英里外的咖啡店，但这并不能诱使我长时间离开会场。会议讨论中充满了各种通过数字技术扩展人类能力的新点子、新方法以及新手段。

据我所知，动物 - 计算机交互会议是一个小得多的活动，可能最多只有 200 名与会者。许多分组议题可能很容易落入其他会议的应用领域，例如增强现实、动物行为或分布式系统。这些论文并没有走在计算机科学的最前沿，但把它们放在一起却使我们回想起农业是如何在数字计算发展早期对这个

领域产生了深远影响，以及计算技术的发展又如何彻底改变了农业的面貌。

我在 ACI 会议上主要看到了三个主题。第一个主题涉及对于动物智能的测试、训练与捕获。与发表在人机交互领域会议上的论文类似，该主题的文章讨论了动物如何对灯光、显示屏幕、振动和其他触觉输出做出反应。

第二个主题则将上述结果运用到现有系统中，以增强动物的工作能力。该主题包含了诸如对工作犬类运动跟踪装置的研究，该装置向人（或中控系统）传送信息并允许其向动物发出命令。这些论文还包括使人们可以与日常呆在家中的宠物进行互动的想法。其中的许多想法已经转化成商业产品，使宠物主人可以监视他们的宠物狗或宠物猫的行为，跟踪它们的运动，用食物奖励它们，甚至和它们玩耍。这些设备可以在房间内投射一个运动的小激光点来逗宠物玩。我知道猫喜欢这样的设备，因为我见过许多宠物猫主人这么做。但我从来没有见过对激光游戏感兴趣的宠物狗，这或许仅仅是因为我没有接触到适合的宠物狗。也许 ACI 的论文可以解释如何让宠物狗对这种玩具感兴趣，或者介绍如何通过其他办法让它们与电脑玩耍。

会议的最后一个主题涉及传感器如何用于监视和管理农场或动物园中的动物，这可以被视为物联网

网和机器学习在动物管理中的应用。例如，一篇论文建议农民可以在牛身上放置传感器以跟踪它们的运动、环境、食物摄入量、步态以及相互作用。根据这些信息，农民可以知道如何更好地管理动物，如何治疗患病动物，以及如何为动物出栏做准备。

最后这个主题直接指向计算与农业的早期联系，以及农业对计算和计算模型如何产生越来越强的依赖。最开始将这两个领域联系在一起的是一名早期的计算机设计师约翰·阿塔纳索夫(John Atanasoff)。阿塔纳索夫是20世纪30年代爱荷华州立大学(Iowa State University)的一位物理学教授，这所学校坐落在远离美国主要工程中心的一座小城。1938年，他设计了一台小巧并有些古怪的计算机，它从未完全正常运行，但最终却引起了美国东部城市制造计算机的研究者的注意。

在计算机史上，阿塔纳索夫很少被人讨论，每当被提及他都被视为一位孤独的天才——他能够预见到某种别人无法理解的技术的重要性，但他却不与其他研究人员接触，独立地进行研究。这种看法完全忽视了阿塔纳索夫的贡献，他那台有点古怪的计算机可以执行一种计算，一种对农业研究和管理学至关重要的计算。他所设计的计算机可以对矩阵进行高斯消元，这是一种在农业统计分析中非常重要的算法，它允许研究人员将庞大的农业数据归约到简单的线性模型。利用这些模型，人们可以确定农作物的最佳种植地点、最佳播种时间以及适用于耕种土地的最佳肥料用量。

高斯消元法并不复杂，但它的过程冗长，需要不断检查。一个人对一个六七行的矩阵进行消元并验证结果的正确性要花大约三个月的时间。而到1950年，电子计算机已经可以在不到一个小时的时间内对300行的矩阵消元。更强的算力促使人们开始以新的方式思考农业。他们不再认为这是一个扎根于传统的领域，转而开始将其视为一种以数据和模型为基础的行业。

农业和计算之间的早期互动是短暂而强烈的。在差不多十年的时间里，期刊上不断发表关于高斯消元新应用的文章。最初这些应用来自农业，很快

便被推广到医疗统计和经济学等领域。这些算法也被应用在工程和物理学中，但它并未像在医学、经济学和农业领域那样产生很大的影响。

我并不是很清楚动物-计算机交互是否也会像农业与计算机在20世纪40年代发生的原始互动那样产生同样的影响。据我所知，该领域的大部分研究涉及的技术在其他领域已经发展得相对比较成熟了。该领域的研究成果可能会增强动物的工作能力，或是对大型动物养殖产生一定影响，不过我相信更简单的技术形式也许会对这些产业产生更大的影响。如果说它改变了什么，也许它最终改变的是我们对于通用智能——即拥有完整智能特性的独立系统的理解方式。75年来我们一直在努力去了解人类的通用智能，间或取得了些许进步。目前我们最好的计算系统已经能够在特定领域展现智能，但将这些领域内取得的成果整合成通用的智能系统却更为困难，在这方面我们目前能做到的事情极为有限。



动物被认为是较简单的生物。通过了解它们与机器的交互方式，我们可以更好地理解智能。与此同时，动物也时常让人们大吃一惊，它们有时拥有难以预料的智慧和洞察力。因此通过研究动物如何与机器交互，我们可以更多地了解智能的外延，以及更好地理解用计算系统捕获智能的困难之处。■

戴维·阿兰·格里尔 (David Alan Grier)

2018CCF杰出贡献奖获得者。电气与电子工程师协会计算机学会(IEEE-CS)前任主席、IEEE Fellow(会士)。乔治·华盛顿大学名誉教授。他目前是华盛顿特区Djaghe LLC公司的技术总监。IEEE Computer 杂志主编。
grier@gwu.edu

CCCF特邀译者：

孙晓明 中国科学院计算技术研究所研究员

人物专访

用“文章”架起交流的桥梁

——专访CCF杰出贡献奖得主戴维·阿兰·格里尔

韩玉琦
中国计算机学会

主编按:从2013年3月起,戴维·阿兰·格里尔(David Alan Grier)教授为《中国计算机学会通讯》(CCCF)撰写专栏文章,迄今已发表66篇。他的文章有故事,有卓见,以小见大,发人深省,受到读者的高度好评。今年中国计算机学会(CCF)授予他CCF杰出贡献奖及荣誉会员称号,以表达CCF及中国计算机界对他的尊敬。

一个外国学者为什么有持续的动力坚持为CCCF写专栏文章?在本期发表的“人物专访”中,格里尔教授阐述了他的人生理念和动力源泉。他认为,“文章”可以架起一座桥梁,交流可以使计算机成为一股为善的力量。他的采访谈话告诉我们,“交流”是需要着力培养的一种基本素质。

2019年1月19日,中国计算机学会在北京举行隆重的颁奖大会。电气与电子工程师协会计算机学会(IEEE-CS)前主席,美国乔治·华盛顿大学名誉教授戴维·阿兰·格里尔荣获CCF杰出贡献奖,并获得CCF荣誉会员称号。在他上台领奖的瞬间,我的眼睛似乎模糊了,为他感动,也为他自豪。

戴维·阿兰·格里尔教授受CCF秘书长杜子德之邀,从2013年3月起为《中国计算机学会通讯》撰写专栏文章,迄今已发表66篇。他的文章范围广泛,以小见大,有深邃的思想,受到CCF会员及其他读者的高度评价,提升了CCCF在业界的影响力。他高度认同CCF的文化并努力帮助CCF发展。他获得这一殊荣完全是实至名归。

从2013年他写的第一篇文章起,我作为CCCF编辑部的主任,便开始和他联系。几年来,虽然只在2013、2014的CNCC和2015的CCF颁奖会上见过寥寥数面,但我们通过邮件往来仿佛每个月都



CCF理事长高文(左)向David颁发CCF杰出贡献奖

在“见面”。久而久之，除了交流文章外，还会谈起些华盛顿的天气，北京的天气，他还会发来在姐姐家农场度假时，拍摄的漂亮的马的照片等。他每个月会准时发来文章，如果偶尔邮件“停摆”，我没能及时回复他，他便会再发邮件询问。像一个时刻被牵挂的亲人。他的每篇文章我都认真阅读，虽然文字不多，但每篇都有放矢，言之有物。他的文章涉猎广泛，不但有“‘云’之得名”“问路”“机器中的‘魅影’”等漂亮的标题，有对物联网、高性能计算、人工智能、敏捷计算、软件工程等发展中遇到问题的分析，还有 IEEE 制定的一些标准介绍，让我们看到了先进学会的一些做法。他的文章还常常提到中国，以拉近中国读者的亲切感。尤为难得的是，他的文章体现出的犀利观点。如 2014 年第 10 期他写《切实负起责任》一文，提出了一个紧迫的问题：“互联网治理很可能发生在未来两年之内，希望 CCF 能够加入到这场变革。”他在 2018 年第 6 期讲计算机科学与计算机工程时说：“计算机科学并不是由技术定义的，它是由机器与人类思考世界的方式之间的联系所定义的。随着计算进入第三代，这一领域仍由这种力量所驱动，无论将来的技术是什么样子。”他于 2018 年第 11 期写的《正确的基础》，提出了一个冯·诺伊曼想解决但来不及解决的深层次问题：计算的神经模型（统计模型）如何与布尔代数的逻辑联系起来（所谓“计算的第二语言”）。他说到：“现在统计人工智能已经在计算机科学领域占据了一席之地。然而，与构成计算科学基础的逻辑学和离散数学的方法完全不同，它似乎代表着另外一种与传统计算机科学完全不同的根基。”等等。他的文章，常常给我启发、震撼，也让我十分钦佩他的好学与知识广博。

2019 年 1 月，当他来到北京，我们再次相见时，犹如一对老朋友，感觉非常亲切。我采访他时，也像久别重逢的知己，那么自然。

问：嗨，戴维，祝贺你获得 CCF 杰出贡献奖。对得到这一奖励，你有什么感想？获奖之后，你还会继续为 CCCF 撰写文章吗？

戴维：谢谢。得到这个奖项，我感到非常荣幸与自豪，感谢推荐我的人，我也非常感谢大家对我文章的认可。得到这个奖励，让我感到写“专栏文章”的意义。是的，只要对 CCF 有好处，能够促进交流，我愿意继续写下去。

问：从 2013 年起，你已经为 CCCF 写了 60 多篇专栏文章，是什么理念和动力使你能坚持多年，愿意为这本刊物写文章？

戴维：我的事业是建立在计算机领域的，但我

总是超越计算机领域来考虑这项技术如何与人类的其他活动相关联。我希望有各种不同的方法来对待技术，并希望人们看到它们之间的联系。我想让这些学科互动并相互支持。我希望以不同方式对待计算的人能够理解使用计算机的所有不同方式。我想通过在 CCCF 上的“文章”架起一座桥梁，努力帮助软件工程师，让中国的工程师和北美、欧洲的工程师进行交流，把他们的工作放在一个更大的背景下去发展，我想尽我所能去帮助他们。这也是我一直致力于此项事业的动力。

问：你的文章，涵盖很多内容，既反映了你的观点，又体现了你的价值取向。如你的第一篇文章《中



2014 年在郑州参加 CNCC 时，
CCF 秘书长杜子德与 David(右)交谈

国比你想象的要近》，你想通过这篇文章，告诉人们什么？

戴维：我想通过《中国比你想象的要近》告诉读者，我们有许多相同之处，这些是我们“对话”的基础。我从2006年起，一直在IEEE-CS的旗舰刊物*Computer*上写文章，关注人类在技术方面的故事，不止是那些巨大的成就，也关注平凡的贡献和作出贡献的个人。我希望通过这些故事去探索计算技术的本质，以及我们为什么会付出那么多精力和时间去热爱这个领域。

我从我父亲那代人获知，计算机将成为一股为善的力量。计算机可帮助我们了解到很多东西，扩展业务，改善制造业，并加强贸易。也许现代人不一定同意这种观点，或者可能很难接受它。现在，计算机已经变得非常普通，成为了人们生活的一部分。而我一直试图用国际化的视角看待计算机事业。我生活在华盛顿特区，倾向于把任何事务都当作国际化的活动。这也是我为什么愿意为CCCF写专栏文章的原因。而在美国，我发现也有很多人想了解关于中国计算机行业的发展现状，他们希望了解中国。

问：作为IEEE-CS前主席，你是怎样看待与CCF的合作，在合作中你都做了哪些工作？你为什么愿意与CCF合作？

戴维：作为IEEE-CS的主席(president)，我2013年来到中国，来到北京，是为合作与友谊而来。这也是我的使命。在北京，我和CCF秘书长杜子德先生每天都见面，我们相谈甚欢，我们讨论了双方可能的合作事宜。因为，我希望探索中国计算机事业的发展如何影响到美国的计算机事业，反之亦然。我也希望讨论IEEE-CS和CCF如何共同促进计算机技术的发展以及如何互相帮助。我想双方所做的贡献会超出以前的想象，我们双方的合作也会比之前所想像的更加紧密。当然，还有一项，就是接受了杜子德的邀请，为CCCF撰写专栏文章。

问：你的成长经历是怎样的？哪些人，哪些事对你的成长产生了大的影响？



2018年6月，David用轮椅推着腰受伤的杜子德（前）访问美国国家历史博物馆计算机馆

戴维：我的父亲对我的成长有巨大的影响。他是一名计算机科学家，是1950~1960年代的计算机设计师，为计算机事业的建立作出过开创性贡献。他创办了宝来计算机公司(Burroughs Computer Corporation)，并在公司负责与客户的关系。他花了很多时间成为公司与客户之间的调解人。他是一名解决问题的“专家”，解决问题是他最好的技能之一，也是让我深深钦佩的。我真的很佩服并欣赏他的能力。

每逢星期六我会去他的办公室，并喜欢“摆弄”计算机，从而了解了计算机。我父亲给了我培训手册，并教我学习编程。当我在中学时，我就编写了可以进行计算和绘制图形的程序。

我在大学和研究生院学习的是数学，数学则教会了我如何通过许多不同的方法去融合(理解)不同种类的技术科目和科学。我上的大学是米德尔伯利学院¹(Middlebury College)，它位于新英格兰佛蒙特州明德镇。这是一所不大的学院，主要培养作家、

律师、商业界人士。它有 200 多年的历史，也是美国最古老的高等学府之一。当我还是一名大学一年级的学生时，我几乎所有的空闲时间都在写一个文字处理程序，因为那时还没有商业用的文字处理程序。我常常去我父亲的办公室，看到过文字处理机，我就决定自己写一个。

当我的教授看到我的论文被印到电脑上时，他们认为是电脑写的论文。我不得不花时间跟他们解释，这篇论文是我写的，电脑只是将论文做了排版格式。当时对大多数教授来说，这是他们第一次看到电脑可以帮助人们写作。

问：你在 CCCF 上发表了那么多的文章，有读者给你写信吗？他们会关心什么问题？

戴维：是的。几乎每期的文章发表后，我都会收到读者写来的信。虽然信并不太多。其中有一封令人印象深刻。我不记得他关注的那篇文章的主题了，我需要查一下才能准确地告诉你你是哪篇文章。我记得读者花了大量时间和精力想让我知道一些事情，去理解他讲的一些问题。我们用“谷歌翻译”交换了两三封邮件，是为了确保我们相互理解而不致有偏差。我发现整个过程相当有趣。他觉得他很有必要联系我，帮助我理解他的观点。我很高兴有这些读者，非常感谢他们对我的关注。我还想鼓励那些在我的文章中看到问题的人和告诉我写些什么内容的人。当下，人们为“翻译”创造了那么多的工具，我和读者之间的交流越来越容易了。如果有读者认为还应该有一些涵盖的主题，欢迎大家写信告诉我，我也许会在未来写到它。

问：许多人认为在写作上没有那么多的话题，你的话题很多，这些内容是怎么确定的？你的文章写得很有意思，在写作上，有什么技巧吗？现在，CCF 会员已有 5.5 万人，意味着你的读者会更多。

你想对他们说些什么？

戴维：我写的都是我感兴趣的计算机领域的课题，我对很多主题都很感兴趣。通常，我会从朋友那里获得主题，或者从我看过的计算机技术或学术期刊的文章中获取主题。我看到一个主题并自问，我是否可以通过它来探索计算机科学的某些问题，然后我会考虑这个话题是否会引起中国读者的兴趣。确定了之后，我就去了解这个主题，并做额外的研究，以获取更多的关于这个主题的信息。

我上的大学是一所要求所有学生都具备语言技巧和写作技能的大学，所以，我在那里做了一些很特别的工作，学习如何写作和如何写得更好。我对如何交流技术也很感兴趣，以及如何对一个好的技术想法的表达。在写文章时，我一直在寻找好的例子和好的想法，以便使我的文章更生动。此外，我还在 IEEE-CS 和 ACM 开设了一门如何写作的课程，我在教导学生如何写作时，自己也收获了很多写作方面的心得。

我认为“好奇心”很重要。我是个好奇的人，想知道世界是如何运作的，虽然我的专业是数学，我也学习了文学，学习了历史，我还读了许多令我感兴趣的各式各样的书籍。

CCF 会员多了，读者多了，是令人高兴的事。提到建议，我认为，当你给别人尤其年轻人提出建议时，要真的提供建议。我的建议有三点：“学习如何学习”、“学习如何交流”及“亲自参与到这项技术中来，使它成为你生活中的一部分”。这三方面，是我的感悟，在我身上的效果很好，我也希望会对别人起到很好的作用。



韩玉琦

CCF 2012~2018 编辑部主任。
yqhan@ccf.org.cn

¹ Middlebury College，亦翻译为明德学院，是全美顶级文理学院之一，位于美国新英格兰北部佛蒙特州明德镇。佛蒙特州明德镇是美国最美丽的山谷之一。明德学院以其卓越的学术声誉在美国文理学院中名列第五，也是美国最古老的高等学府之一。《美国新闻与世界报道》(U.S. News and World Report) 将其分类为“最为挑剔”的学校。

CCF 推荐 B 类国际学术会议介绍

2018自然语言处理实证方法 会议概览

关键词：自然语言处理 计算语义学

岂凡超 韩旭 刘知远
清华大学

大会概况

2018 自然语言处理实证方法会议 (2018 Conference on Empirical Methods in Natural Language Processing, EMNLP2018) 于 2018 年 10 月 31 日至 11 月 4 日在比利时布鲁塞尔广场会议中心召开。EMNLP 是自然语言处理领域三大顶级会议之一，由国际计算语言学协会 (Association for Computational Linguistics, ACL) 下属的 SIGDAT 专委会承办。

本次 EMNLP 参会人数、投稿量等都达历史之最。参会人数超过 2500 人，比去年增加了一倍多。大会收到 2231 篇有效投稿，比去年增加 46%，其中录用 549 篇，录用率为 24.6%，比前三年 26% 的录用率略有下降。

本次会议共安排了 600 多场丰富的活动，分别

在不同会场同步进行，并各具特色，满足了不同参会者的需求。

主旨报告

本次会议有三场主旨报告。荷兰格罗宁根大学 (University of Groningen) 的约翰·博斯 (Johan Bos) 教授作了题为 “The Moment of Meaning and the Future of Computational Semantics” 的报告，讨论了计算语义学将在未来的自然语言处理研究和应用中发挥的作用，并且提出采用语言中立的意义表示来进行推论。来自美国哥伦比亚大学 (Columbia University) 的朱莉娅·赫希伯格 (Julia Hirschberg) 教授作了题为 “Truth or Lie? Spoken Indicators of Deception in Speech” 的报告，介绍了关于谎言检测的研究工



作。她的团队利用声学、韵律、人口统计等特征在大规模语音语料库上训练分类器，实现了更准确的谎言检测。她还进一步研究了性别、个性和母语等特征对说谎行为的影响。第三场报告的题目是“Understanding the News that Moves Markets”，报告人是美国彭博公司(Bloomberg L.P.)的数据科学部门主管吉迪恩·曼(Gideon Mann)博士。他在报告中回顾了语言技术如何帮助市场参与者迅速了解重大突发事件并做出响应，并介绍了目前应用于金融领域的最前沿的自然语言处理技术。

最佳论文

本次会议评选出了2篇最佳长论文、1篇最佳短论文和1篇最佳资源论文。第1篇最佳长论文题目为“Linguistically-Informed Self-Attention for Semantic Role Labeling”。该文提出了基于语言学的自注意力模型，将multi-head自注意力机制与包含依存分析、词性标注、谓词检测和语义角色标注在内的多任务学习相结合，实现了在语义角色标注等多个任务上较大的性能提升。第2篇最佳长论文是Facebook的纪尧姆·兰普尔(Guillaume Lample)等人的“Phrase-Based & Neural Unsupervised Machine Translation”。该文研究了如何只利用大规模单语语料实现无监督机器翻译，并提出了基于神经网络和基于短语的翻译模型。

最佳短论文是卡内基梅隆大学团队的“How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks”。该论文测试了现有神经网络模型在bAbI、SQuAD等流行的阅读理解数据集上的性能，发现仅需要问题信息或文章段落信息的模型在上述数据集中已经有较好的表现。

来自剑桥大学团队的“MultiWOZ—A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling”获得了最佳资源论文，该论文构建了一个面向任务型对话建模的大规模跨领域数据集。

会议亮点

新任务与新数据集的发布。在过去，绝大多数论文都是在已有问题的已有数据集上提出新方法、新模型，而这些基准数据集往往十分陈旧，要么问题已经被基本攻克，要么与实际应用场景有较大脱节。本次EMNLP录用了许多数据集论文，其中一部分在论文中也提出了新的任务，例如问答领域的“A Dataset and Baselines for Sequential Open-domain Question Answering”和“HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering”，关系分类领域的“FewRel: A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation”，对话领域的“A Dataset for Document Grounded Conversations”，等等。这些新数据集要么更加贴近自然语言处理应用实际需求，要么提出了自然语言处理面临的新挑战，如少次学习(few-shot learning)等，基本反映了自然语言处理研究的最新趋势，对推动自然语言处理领域研究和应用的发展意义重大。

大规模数据的充分应用。EMNLP一直聚焦自然语言处理中的经验方法(empirical methods)，数据驱动的深度学习技术正是经验方法的典范。本次EMNLP也充分展现了自然语言处理领域对预训练更加重视的趋势。以ELMo、BERT为代表的基于大规模文本数据的预训练模型，在自动问答、阅读理解、情感分析、机器翻译等任务上均表现出显著的性能优越性，已经吸引了研究者的广泛关注。在本次会议上，有相当一部分工作在ELMo等预训练模型的基础上对模型进行精调，从而获得更佳性能。本次会议最佳长论文之一“Phrase-Based & Neural Unsupervised Machine Translation”，旨在充分利用规模更大的单语文本数据，尝试构建无监督机器翻译模型。进入大数据时代，如何更好地利用大规模文本数据提升自然语言处理效果，将成为自然语言处理领域持续关注的重要研究方向。

来自工业界的深度参与。近年来，在几乎所

(下转 73 页)

梦想在 CCF 实现

——CCF 秘书处总结及表彰年度优秀

2018 年是 CCF 发展进程中极不平凡的一年，为了更好地对年度工作进行回顾和总结，也为了找出工作中存在的问题，继续打造一支更具战斗力的队伍，2019 年 1 月 22~23 日，CCF 秘书处在京召开了 2018 年度总结及表彰会议。

CCF 秘书长杜子德首先对 2018 年 CCF 的各项工作予以肯定，“2018 年 CCF 的发展达到了一个新高度，在这一年里各项工作不断地取得新业绩、新成果，CCF 会员人数突破 5.5 万人，CNCC 参会人数突破 7000 人，会议系统 2.0、数字图书馆二期顺利上线，CSP 年度报名人数突破 2.5 万人次，NOI 教师培训达 13 次”，同时也表示“一项项成绩的背后，除了每一位 CCF 会员为 CCF 发展作出重要贡献外，也包含了 CCF 秘书处每一位专职员工的辛勤付出。正是由于他们努力拼搏、恪守职责才使 CCF 的发展达到了一个新高度。CCF 秘书处是一支能征善战的优秀队伍，他们也是 CCF 最宝贵的资源之一。感谢每一位员工为 CCF 发展付出的极大努力。”

在总结会上，秘书处的员工分享了 2018 年经历的艰辛和取得的成绩，并表示“感谢 CCF 为我们提供了实现梦想的舞台，2018 年难忘的回忆成为历史，我们将满怀希望与憧憬迎接 2019 年新的征程”。本次会议还特别邀请了 CCF 会士、常务理事、中国人民大学教授杜小勇，CCF 会士、抗恶劣专委主任刘爱民研究员，CCF 理事、微软亚洲研究院学术合作总监马歆女士，CCF YOCSEF 总部学术秘书、滴滴出行科技合作总监吴国斌博士，CCF 特邀摄影师赵明理先生出席。嘉宾们在听取秘书处员工分享后，肯定了秘书处的工作，并鼓励大家要激情投入，为会员提供更加专业化的服务，打造出更多具有影响力的产品，使 CCF 从优秀走向卓越。



2018 CCF 颁奖大会上，CCF 高层领导、秘书处员工和嘉宾合影

在表彰环节，对 2018 年度表现优异、作出重要贡献的个人或团队颁发了秘书长特别奖、优秀员工奖、CCCF 刊物发展奖、优秀新员工奖、优秀项目、优秀团队等六个奖项。鼓励全体员工向他们学习，站在新的起点上，正视前进中面临的困难与挑战，以更加坚定的信念、更加饱满的热情、更加务实的作风、更加强大的合力，为 CCF 的发展继续贡献力量。

■ 秘书长特别奖

张建泉

■ 优秀员工奖

富 蕾 马 琳 周 苗

■ CCCF 刊物发展奖

韩玉琦

■ 优秀新员工奖

郑智盛 王亚松 白素勇 刘 霞 郑 新

■ 优秀项目奖

NOI 项目

■ 优秀团队奖

会员部

(上接 71 页)

有人工智能领域的学术会议上，工业界参与学术研究并发表高质量论文已经稀松平常，企业与高校联合进行研究并发表论文也日益常见。本次 EMNLP 也不例外，录用了很多来自 Google 等企业的研究论文，并且这些工作在口头报告和海报环节都备受关注。由于企业拥有更庞大的科研团队和更高效的内部协作，拥有直接来自商业产品的数据资源，以及更强大的计算与存储资源，工业界从事的很多研究工作是高校难以开展的。高校与企业在学术研究方面具有高度的互补性。因此，可以预见未来高校与企业的学术合作会更加密切，学术研究将不再是高校和科研院所的“专利”。而从企业参与 EMNLP 的热情程度可以看到，近年来工业界开始日益重视自然语言处理的研究应用价值。美国彭博公司近年来持续站台 EMNLP，也表明自然语言处理技术已经在新闻、法律、金融等领域生根发芽，并得到相当程度的应用认可。

挑战

自然语言是人类交流信息、表达情感、开展工作的重要工具，自然语言处理是实现人工智能的关键，有着重要的科学意义和应用价值。自然语言处理技术与应用正受到越来越多研究者和企业的关

注。然而，我们也应清醒地意识到，自然语言处理距离彻底实现让计算机理解语言，还面临很多挑战。例如，如何在深度学习中考虑语言知识、世界知识、常识知识；如何在复杂多模态语境下实现语言理解；如何在开放对话环境中实现对复杂上下文的理解与建模；如何充分利用互联网中的大规模数据，等等。期待有更多有识之士参与进来，共同探索自然语言处理这个充满未知和无限可能的领域。 ■



■ 岚凡超

清华大学博士生。主要研究方向为自然语言处理。qfc17@mails.tsinghua.edu.cn



■ 韩 旭

清华大学博士生。主要研究方向为自然语言处理。hanxu17@mails.tsinghua.edu.cn



■ 刘知远

CCF 高级会员、CCCF 编委。清华大学副教授。主要研究方向为自然语言处理、知识图谱。
liuzy@tsinghua.edu.cn

新技术 & 新应用

具有指数加速可能的量子机器学习算法被提出

清华大学量子信息中心段路明团队发现了一种具有指数加速可能的量子机器学习算法，该成果的研究论文“*A quantum machine learning algorithm based on generative models*”于2018年12月7日发表在*Science Advances*。

他们提出了一种基于优化多体量子纠缠态的量子生成模型，并证明了该量子生成模型在学习能力与预测能力方面都存在指数加速。在量子生成模型

中，经典图中表示概率的参数由正实数扩大至复数域，这种新的量子图模型所需的参数个数相比于经典图模型有指数级的减少，这对于生成模型来说，在空间和时间的效率上都是巨大的优势。基于这一模型，他们进而提出了启发式量子机器学习算法。该算法可以将生成模型的推断和训练问题转化成量子多体哈密顿量的基态制备问题，并由此证明量子算法的指数加速。

AI 在医疗方面取得革命性突破

1月7日，吴恩达团队在*Nature Medicine* 上发表了论文“*Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network*”，该研究利用深度神经网络仅通过心率数据就可以诊断患者是否心律失常。心电图数据全部由专家标注，分成12种不同情况，包括10种心律失常，窦性心律以及噪音。基于这些数据，研究人员训练了一个包含33个卷积层和线性输出层的神经网络。只需输入心电图数据，该系统就可以将其拆分成时长1.28秒的数据样本，判断每个1.28秒属于12种心率（及噪音）中的哪一种。实验表明其准确度高达83.7%，超过了人类心脏病医生。

同期，*Nature Medicine* 上发表了论文“*Identifying facial phenotypes of genetic disorders using deep learning*”。该研究使用了17000多张患者的面部图像来训练神经网络，所有患者被确诊的遗传综合征总计超过200种。研究人员利用两个独立的测试数据集测试该深度学习算法DeepGestalt的表现，每一个数据集都包含数百张之前经过临床专家分析的患者面部图像。对于每一张测试图像，DeepGestalt被要求列出每张面部图像可能代表的综合征，并按不同综合征的概率依次排序。在两组测试中，成功率达到90%左右，DeepGestalt提出的前10个答案中都包括了正确的综合征。研究结果表明人工智能有望在临床实践中，辅助罕见遗传综合征的优先级划分与诊断。

微软公开小冰系统设计

2018年12月21日，微软小冰团队发表论文“*The design and implementation of XiaoIce, an empathetic social chatbot*”，公开了其聊天机器人“小冰”的开发过程，详述了设计原则、系统架构和关键组件，展示了小冰是如何在长时间的对话中动态地识别人类的情感和状态，理解用户的意图，并响应用户的需求。

研究团队在文中强调，作为一个独特的人工智

能伴侣，小冰可以在情感上与人们产生联系，满足人们对交流、感情和社会归属的需求。在小冰的系统设计中，智商与情商的结合是核心，同时又具有独一无二的个性。针对这样的设计原则，他们将人机社交聊天视为基于马尔可夫决策过程(MDP)的决策，并针对长期用户参与度和每次会话的对话轮数(Conversation-turns Per Session, CPS)进行优化。

Transformer 再升级

Transformer 是谷歌在 2017 年提出的 NLP 框架，在机器翻译领域，它已经几乎全面取代递归神经网络。2019 年 1 月，来自卡内基梅隆大学和谷歌的研究人员提出了 Transformer 的升级版：Transformer-XL。

Transformer-XL 可以使 Transformer 能够在不破坏时间一致性的情况下学习超过固定长度的依赖关系。Transformer-XL 包含一个 segment-level 的递归机制和一种新的位置编码方案。这样不仅可以捕获长期依赖关系，还可以解决上下文碎片问题。实验

表明，Transformer-XL 学习的依赖关系比 RNN 长 80%，比 vanilla Transformer 长 450%，在短序列和长序列上都能获得更好的性能。并且在评估过程中比 vanilla Transformer 快 1800 多倍。此外，这一新架构在 5 个数据集上都获得了强大的结果。1 月 18 日，相关论文“Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context”放在 arXiv 的版本更新了更好的结果，并且公开了 Tensorflow 和 PyTorch 版本的代码、预训练模型和超参数。

欧盟将投入 200 亿欧元发展人工智能

2018 年 12 月，欧盟宣布计划在 2020 年底，实现私人投资和公共投资至少 200 亿欧元，用于发展人工智能。经过近六个月的会议讨论，欧盟成员国就以下内容达成一致。(1) 通过合作伙伴关系将投资效益最大化。投资目标包括：所有成员国将制定自己国家的人工智能发展战略；建立新的人工智能研究和创新合作伙伴关系；建立新的人工智能扩大基金，为初创公司提供支持；开发和连接欧洲人工智能卓越中心。(2) 创建欧洲数据空间。致力于开发人

工智能技术需要用到的大型、安全且可靠的数据集。(3) 培养人才、技能和终身学习精神。将通过专门的奖学金等方式支持人工智能领域高等教育，充分利用蓝卡(Blue Card)系统留住和吸引欧洲的高技能人工智能专业人士。(4) 开发有道德且值得信赖的人工智能。专家小组目前正在制定开发和使用人工智能的道德准则，让欧洲注重道德的方法进入全球舞台。

欧盟委员会希望通过实施其人工智能战略，促进人工智能在欧洲的开发和使用。

IEEE-CS 公布 2019 年十大技术趋势

2018 年 12 月 18 日，IEEE-CS 公布了 2019 年的技术发展趋势：(1) 深度学习加速器，如 GPU、FPGA 和 TPU；(2) 辅助驾驶，该技术高度依赖于深度学习加速器进行视频识别；(3) 身联网 (Internet of Bodies, IoB)，物联网和自我监测技术正在更加靠近人体甚至进入人体内部；(4) 社会信用算法，通过生物识别技术和混合型社交数据流的结合，可以将观察转化为对个人的好坏及是否值得得到公众社会认可的判断；(5) 先进（智能）材料和设备，他们将在医疗保健、包装、家电等领域创造激动人心的应用；(6) 主动安全保护，如在新的攻击类型暴露时能被激活

的钩子以及识别复杂攻击的机器学习机制等；(7) 虚拟现实(VR)和增强现实(AR)，除了游戏方面，VR 和 AR 技术在教育、工程和其他领域也可以发挥巨大的作用；(8) 聊天机器人，除了基本客户服务、虚拟助理，业界也在寻求将这一技术作为提供治疗支持的一种方式，例如扩展到与认知障碍儿童的互动；(9) 自动语音垃圾(robocall)预防，这一技术现在可以阻止被假冒的呼叫者 ID，并拦截可疑来电；(10) 人性化技术(特别是机器学习)，机器学习、机器人和无人机的大规模使用将有助于改善各行各业的工作流程和工作效率。

(本栏目内容由动态栏目编委鲍捷提供整理)

机器学习如何影响本科生 计算机课程 *

作者: 本杰明·夏皮罗 (R. Benjamin Shapiro)

丽贝卡·菲布林克 (Rebecca Fiebrink)

彼得·诺维格 (Peter Norvig)

译者: 刘如意 史媛媛 苗启广

关键词: 机器学习 计算学科教育 本科生

机器学习日益增长的重要性给计算机科学教育带来了挑战性的问题。

现在, 机器学习被广泛应用于大多数领域, 从语音识别系统到搜索引擎、自动驾驶汽车及监狱判刑系统。许多曾经由人类设计和编写的应用程序已经将人工编写的组件与从数据中学习到的行为相结合。这种转变给计算机科学 (Computer Science, CS) 从业者和教育者带来了新的挑战。在本专栏中, 我们考虑机器学习如何改变我们认为的计算机科学的核心知识和技能, 以及这将如何影响机器学习课程和更广泛的计算机科学大学课程的设计。

不要数学家思维, 而要像科学家那样思考

计算机教育工作者^[1,6]历来认为: 计算机科学的核心是用数据结构和算法表示的人类可理解的抽象的集合。确定的和逻辑上可验证的算法一直是计算机科学知识论和实践的中心。

随着机器学习 (Machine Learning, ML) 的发展,



这种情况发生了变化: 首先, 典型的模型可能是由数百万参数组成的不透明组合, 而不是人类可读的算法。第二, 验证过程不是正确性的逻辑证明, 而是有效性的统计证明。正如兰利 (Langley)^[5] 所观察到的, 机器学习是一门经验科学, 它与物理和化学

*本文译自 *Communications of the ACM*, “How Machine Learning Impacts the Undergraduate Computing Curriculum”, 2018, 61(11): 27~29 一文。

等领域共享知识论方法。

传统软件是由描述完成目标所需步骤（如何实现）的人类程序员构建的，而典型的机器学习系统是通过描述系统尝试最大化的目标（实现什么）来构建的。学习过程使用样本集来确定实现目标最大化的模型。训练好的模型扮演着数据结构和算法的角色。每个参数所扮演的角色对人类来说并不清楚，这些计算解决方案不再反映人类对问题域的概念性描述，而是作为数据概括的函数，这些数据概括只能根据它们的经验可测度性能来理解。

为了使机器学习取得好的效果，许多学生将不再专注于算法开发，而专注于数据采集、数据清理、模型选择和统计测试。

计算机科学教育中的机器学习教育

机器学习历来是计算机科学的一个专业领域，但现在从计算机体系结构到操作系统^[3]与核心计算机科学学科越来越相关。甚至可以公平地说，现在机器学习是计算机科学的一个核心领域，它为用于定义和推理计算系统的 λ 演算提供了平行的理论基础。因此，机器学习日益增加的重要性为计算机科学教育带来了具有挑战性的问题：如何将机器学习的实践和理论主题整合到本科课程中？如何在保持本科学位培养计划 (undergraduate degree programs) 总时长相对固定的前提下，以增强而不是取代经典计算机科学技能的方式，为扩充的机器学习内容留出空间？

对入门过程的改变。大多数计算机科学本科培养计划是从强调编程技能发展的入门课程开始的，涵盖诸如控制结构、函数的定义和使用、基本数据类型以及简单算法的设计与实现等内容^[4]。

在许多情况下，这些课程中的作业是利用现有的库函数，例如读写数据到文件系统中。学生不需要完全理解这些库和底层硬件是如何工作的，甚至不需要使用这些库提供的接口。入门课程的目标是培养学生开发用于推理计算机如何执行程序的概念

机^[2]，以及编写和调试计算机可执行程序的实用技能。

这两个目标也可以描述以机器学习为核心的领域的入门课程。我们不认为机器学习方法会取代这类课程中的符号编程，但它们为学生程序中函数行为的定义和调试提供了可供替代的方法。学生将在早期学习两种概念机——经典逻辑计算机和统计模型。他们将学习为每种概念机编写、测试和调试程序的方法，并学习在软件系统中组合这两种模型。

我们设想未来的入门课程将通过利用对初学者友好的程序编辑器、库和作业将机器学习包含进来，这些程序编辑器、库和作业鼓励学生用机器学习来定义一些函数，然后将这些函数集成到使用更传统方法编写的程序中。例如，学生可以拿出他们以前使用经典编程技术创建的游戏作业，然后使用机器学习技术创建一个手势界面（例如，使用智能手机的加速计，网络摄像头的姿势信息或麦克风的音频信息）将游戏中的玩家角色向上、下、左、右移动。这样的作业将鼓励学生参与创建或组织训练样例，度量他们训练模型的性能，以及通过调整训练数据或对学习算法和特征的选择来调试模型。

这些活动不需要深入理解机器学习算法，就像使用高级 API 从文件系统读取文件不需要深入了解计算机硬件或操作系统一样。然而，这些活动可以将计算机科学专业的新生吸引到机器学习的核心认识论实践中，为在其他背景下再次遇到机器学习奠定了基础（无论是机器学习理论类选修课，还是计算机视觉、体系结构或专业软件开发高级选修课）。这样的活动还能够创建新的、吸引人的软件类型（例如，由实时传感器或社交媒体数据驱动的系统），对于新手程序员（甚至是专家）而言，这些软件类型在没有机器学习的情况下很难创建。

对高级课程核心的改变。在大多数计算机科学学位培养计划中，入门课程的学习之后是一系列更高级的课程学习。那么考虑到机器学习，高级课程的核心应该怎样变化呢？

目前软件检验和认证的课程强调两点：正确性验证和对验证程序布尔属性的测试。但在机器学习应用中，其重点在于实验设计和实验结果的统计推

断。未来的课程应该包括数据驱动的软件测试方法，例如测试套件的开发。当使用特定数据训练时，此测试套件可以评估软件工具的执行是否可接受，并且可以监控随时间变化的可测度回归 (measurable regressions)。

人机交互 (HCI) 课程可能会被扩展延伸，用于反映机器学习如何改变可创建的面向人类技术的性质以及这些技术被创建和评估的过程。例如：机器学习能够创建可以动态适应其使用数据的应用程序。人机交互教育目前强调利用心理学和人类学的经验方法来理解用户的需求和评估新技术；现在，将机器学习应用于记录用户与产品交互的日志数据的能力可以推动了解用户体验并将其转化为设计建议的新方法。未来的人机交互课程将需要包括这些基于机器学习的系统设计和评估方法。

操作系统课程为内存分配和过程调度等任务描述了最佳实践。一般来说，这些任务的关键参数值是根据经验来进行选择的。但如果使用机器学习，可以允许参数值甚至整个方法根据实际运行的任务而改变，从而使系统更有效，更适应于变化的工作负载，甚至是设计人员无法预见的工作负载。为了动态优化系统性能，未来的操作系统课程可能需要包括机器学习技术^[3]。

对先修课程和并发期望的改变。计算机科学的课程体系 (CS curricula) 通常要求包括计算机科学系以外的课程，如数学和物理课程。许多情况下，特别是在工科院校制定计算机科学培养计划 (CS programs) 时，要求强调微积分课程。许多培养计划里还包括概率和统计学方面的课程，不过值得注意的是：2013 ACM-IEEE 联合计算机课程体系 (ACM and IEEE's joint Computing Curricula 2013) 的作者们“认为没有必要为所有计算机科学培养计划里的所有专业都开设概率论方面的完整课程。”^[4]

这些建议仍然适用吗？许多培养计划需要概率和统计学方面的课程，这是我们热心鼓励的，因为它们对理解机器学习算法设计和分析背后的理论以及有效地使用机器学习中某些强有力的方法来说是至关重要的。和优化相关知识一样，线性代数对于

机器学习从业者和研究人员来说也是必不可少的。因此，机器学习的基本知识既广泛，又区别于取得计算机科学学位的传统要求。那么，培养未来计算机科学家的必要条件应该是什么呢？

结论

2013 ACM-IEEE 计算机科学课程体系^[4]确定了 18 个不同的知识领域 (KAs)，包括算法和复杂性，体系结构和组织，离散结构和智能系统。对 KAs 的定义和建议的关注时限反映了计算机科学的经典观点，机器学习仅在一些建议的选修课程中提及。我们认为，在过去的几年中，计算机科学中机器学习应用的迅速崛起表明需要重新考虑这样的指导性文件以及计算机系教育课程的相应变化。

此外，关于人们如何学习机器学习的研究是迫切需要的。几乎所有已发表的计算学科教育文献都是关于经典的计算方法。正如我们在本专栏前面提到的，机器学习系统与传统的数据结构和算法从根本上不同，因此必须以不同的方式进行推理和学习。来自于数学和统计教育研究的许多见解可能与机器学习教育研究相关，但这些领域的研究人员很少与计算学科教育研究人员产生交互。因此，我们呼吁资助机构和像 ACM 这样的专业社团，利用其号召力将计算学科教育研究人员和数学学科教育研究人员聚集在一起，支持开发关于机器学习教学和学习的丰富知识库。 ■

作 者：

本杰明·夏皮罗 (R. Benjamin Shapiro)

美国科罗拉多大学教育学院及信息科学系助理教授，计算机科学系 ATLAS 研究所助理教授。
ben.shapiro@colorado.edu

丽贝卡·菲布林克 (Rebecca Fiebrink)

伦敦大学计算系高级讲师。
r.fiebrink@gold.ac.uk

彼得·诺维格 (Peter Norvig)

谷歌公司研究总监。
pnorvig@google.com

读者评论及作者回复

迈赫兰·萨哈米 (Mehran Sahami) (2018年10月30日12:51)

我为这篇文章的作者鼓掌。虽然我完全同意作者的总体观点，但我想指出一些2013 ACM-IEEE计算机科学课程体系(CS2013)中关于概率和机器学习讨论的不准确描述。

例如，作者写到“值得注意的是：CS2013的作者们‘认为没有必要为所有计算机科学培养计划里的所有专业都开设概率论方面的完整课程。’”查看引号中句子所出自的完整句子是有启发性的，“同样的，虽然我们注意到一个越来越明显的趋势——概率论和统计学在计算学科上的运用（这个趋势反映在知识体系中关于这些主题的核心课时数的增加），并且相信这一趋势很可能在未来持续下去，但是我们仍然认为没有必要为所有计算机科学培养计划里的所有专业都开设概率论方面的完整课程。”

引号中的话的重点并不是强调概率论对于计算机科学专业不重要，正如上文作者们在文章中所建议的那样，恰恰相反，概率论的重要性与日俱增并将继续持续下去。尽管如此，在2013年（或许直到今天），并不是计算机科学的所有专业都需要一门完整的概率论课程，尤其是在那些对本科专业所需课程数量有严格限制的学院。例如，一些计算机科学专业的学生（也许那些在培养计划中不强调人工智能的学生可以将概率论作为离散数学课程的一部分来考虑），而其他学生（那些专注于人工智能的学生）将被要求修满（或更多的）概率论课。值得注意的是，在CS2013中，离散结构知识领域中包含8个完整的概率论核心课时。CS2013还包含一些“示范课程”，展示了包括一部分概率内容的“离散结构/数学”以及全部“计算机科学家的概率论”课程，说明了这两种模型在本科课程中的可能性。

此外，作者还提到在CS2013中，“机器学习只在少数建议的选修课程中被提及。”这是不正确的。在“智能系统”知识领域，有两个核心课时的“基础机器学习”。虽然我们完全承认两个核心课时并不多，但将它们包括进来表明在2013年我们就已经意识到计算机科学专业的学生应该接触一些机器学习课程。此外，包括一个（公认是选修的）“高级机器学习”知识单元意在强调，对于专注于人工智能相关领域的学生来说，他们确实应该学习更多机器学习课程。正如贯穿CS2013所提到的，核心课时是学生应该满足的“最低”要求，大多数培养计划包括许多超出核心课时的课程，从而形成一个完整的课程体系。对于从事任何与人工智能相关工作的学生来说，他们当然应该有更多的机会而不仅仅是依靠特定的核心课时来学习概率论和机器学习。事实上，为了达到这个目的，CS2013中有6个范例课程展示了涵盖智能系统领域的不同模型，所有这些模型都包括了核心课时要求以外的很多课程。

即使那样，我重申：上文作者在文章中的观点是言之有理的。机器学习是一个将被继续提高重要性的领域，更多内容被纳入本科课程体系将有利于计算机科学培养计划。CS2013在五年前就已经试图突出这一趋势，它将机器学习的核心课时包括了以前的课程体系指南中不存在的部分，并为以前不存在的高级机器学习创建了一个更全面的选修领域，还包括用来说明如何将这些内容纳入计算机科学课程实例的课程范例。

本杰明·夏皮罗 (2018年10月31日05:59)

亲爱的迈赫兰：

感谢你周到细致的回复。我们承认你所提的机器学习是课程体系核心的观点是有效的。而且，我

们同意一些学校确实给机器学习提供了良好的覆盖范围，但请注意，ACM-IEEE 核心课程体系并不要求它们这样做。

我们的论点的本质是，机器学习不再是计算机科学中的一个边缘话题，而是转移成了新的计算机科学家需要知道的核心内容。从这个角度来看，在 ACM-IEEE 联合课程推荐中，将机器学习作为选修课的处理目前是不合适的。虽然范例选修课的描述对于说明各院系如何纳入机器学习是有用的，但如果他们选择这些课，这些范例课仍然是选修课，而不是课程体系的核心要求。我们希望未来计算机课程体系的修订把机器学习作为核心，同时对概率和统计教育的建议进行相应的修改。

CS2013 将其内容要求和建议分为三个部分：核心一级 (Core Tier-1)、核心二级 (Core Tier-2) 和选修课 (Elective)。CS2013 第 29 页将这些术语描述如下：“计算机科学课程体系应该涵盖所有的核心一级主题，所有或几乎所有的核心二级主题，并在许多选修主题中具有明显的深度（例如：对于计算机科学的本科学位来说，光有核心主题是不够的）。随后，CS2013 还说：“核心二级主题通常对于一个计算机科学本科学历来说是必需的。要求它们中的绝大多数是一个最低期望，如果一个培养计划覆盖所有二级主题，我们鼓励他们这样做。计算机科学课程体系的目标应该是覆盖 90%~100% 的核心二级主题，至少也要达到 80%。”

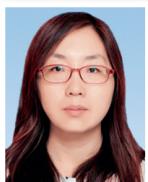
核心一级主题和核心二级主题的本质区别在于：二级列出的主题是强烈推荐的，但不是必需的。换句话说，CS2013 有两类选修课：强烈推荐的选修课和推荐的选修课。所有智能系统的核心内容都在核心二级中，在 ACM 推荐的 308 课时的核心课程中占了 2 课时，并且院系可以选择采用。

一个院系可以根本不选择智能系统的课程，或者选择只排除机器学习部分，并且仍然满足 ACM 课程体系要求。因此，我们认为我们所提的机器学习是 CS2013 中的选修内容的说法是准确的。

致以亲切的问候。

本杰明·夏皮罗
丽贝卡·菲布林克
彼得·诺维格

译者：



刘如意

CCF 专业会员。西安电子科技大学讲师。
主要研究方向为计算机视觉、深度学习等。
ruyiliu@xidian.edu.cn



苗启广

CCF 理事、CCCF 编委。西安电子科技大学计算机科学与技术学院副院长、教授。
主要研究方向为计算机视觉、深度学习、
大数据分析等。
qgmiao@mail.xidian.edu.cn

(本期译文责任编辑：苗启广)



史媛媛

西安电子科技大学硕士研究生。主要研究
方向为计算机视觉等。
sy960315@163.com

参考文献

- [1] Aho A V. Computation and computational thinking[J].
The Computer Journal, 2012, 55(7): 832-835.

更多参考文献：<http://dl.ccf.org.cn/cccf/list>



中国计算机学会青年计算机科技论坛
CCF Young Computer Scientists & Engineers Forum

CCF YOCSEF

激 情

责 任

制 度

YOCSEF 是 CCF 1998 年创建的系列性活动。它以“承担社会责任、提升成员能力”为宗旨，由社会各界有责任、有激情、有思想的青年学者、企业家策划与组织。活动形式有论坛、学术报告会等，每年活动逾 200 场。

2018 CCF颁奖大会在京举行

2018 CCF 颁奖大会于 2019 年 1 月 19 日在北京金隅喜来登酒店举行。颁奖会以“责任·创新·奉献”为主题，颁发了 2018 年度 CCF 终身成就奖、CCF 夏培肃奖、CCF 杰出贡献奖、CCF 卓越服务奖、CCF 杰出教育奖、CCF 计算机企业家奖、CCF 杰出工程师奖、CCF 优秀博士学位论文奖等 8 个奖项。300 余位嘉宾汇聚一堂，见证了这辉煌荣耀的时刻。

颁奖大会由 CCF 秘书长杜子德主持，CCF 奖励委员会主席郑纬民致辞。郑纬民提到：“CCF 奖励除了科技奖外都实行推荐制，这是 CCF 奖励的独特之处，也符合国际评奖规范。CCF 评出来的奖项有较高的公信力，影响力也逐年提升。CCF 还和国际知名学术组织联合设奖，使 CCF 奖项走向国际，扩大在国际上的影响力。表彰优秀，以让更多的同仁为计算技术的发展和应用作出贡献，这是学术组织的重要职能，也希望更多的同仁为我们的社团服务。”

“CCF 终身成就奖”设立于 2010 年，授予 70 岁以上，在计算机科学、技术和工程领域取得重大突破，成就卓越、贡献巨大的资深中国计算机科技工作者。中国人民解放军军事科学院系统工程研究院系统总体研究所研究员、中国工程院院士何新贵和中国科学院软件研究所研究员、中国科学院院士周巢尘获得 2018 “CCF 终身成就奖”。该奖得到了腾讯公司的赞助，CCF 名誉理事长李国杰和 CCF 副理事长、腾讯公司副总裁王巨宏女士为两位获奖人颁奖。全体参会嘉宾起立，用持久的雷鸣般的掌声表达对老一辈科学家的敬意。

2014 年设立的“CCF 夏培肃奖”，授予在学术、工程、教育及产业等领域，为推动中国的计算机事业作出杰出贡献、取得突出成就的资深女性科技工作者。该奖以我国著名女计算机科学家夏培肃先生命名。该奖得到了曙光信息产业（北京）有限



CCF秘书长杜子德主持大会



CCF奖励委员会主席郑纬民致辞

公司的资助。在并行算法研究方面作出杰出贡献的航天工程大学教授李晓梅和在藏文信息处理研究方面作出杰出贡献的西北民族大学教授于洪志荣获 2018 “CCF 夏培肃奖”。CCF 理事长高文和曙光公司高级副总裁沙超群先生为获奖者颁发了奖杯和证书。

“CCF 杰出贡献奖”设立于 2010 年，表彰在某方面对 CCF 发展有独特或重大贡献的个人或单位。IEEE-CS 前主席、乔治华盛顿大学名誉教授 David Alan Grier 和北京大学教授肖建国获此殊荣。该奖得到了阿里巴巴集团的赞助，CCF 理事长高文和阿



何新贵获CCF终身成就奖



周巢尘获CCF终身成就奖



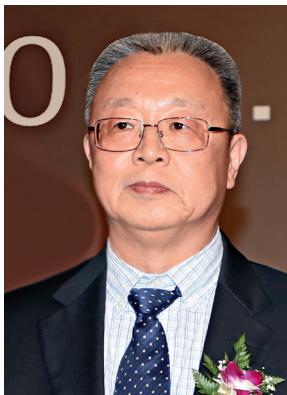
李晓梅获CCF夏培肃奖



于洪志获CCF夏培肃奖



David获CCF杰出贡献奖



肖建国获CCF杰出贡献奖

里巴巴云智能事业群战略与合作部资深总监、达摩院院长助理刘湘雯女士为获奖者颁奖。

同时，为了表彰 David 多年来持续为《中国计算机学会通讯》(CCCF) 撰写专栏文章而为 CCF 作出的贡献，CCF 授予 David Alan Grier 荣誉会员，

CCF 理事长高文为他颁发了 CCF 荣誉会员证书。

“CCF 卓越服务奖”设立于 2011 年，授予为 CCF 连续服务十年以上并有重要贡献的会员。广州科韵大数据技术有限公司创始人、拓尔思知识图谱研究院院长臧根林博士获得 2018 “CCF 卓越服务奖”。阿里巴巴集团赞助了该奖项，CCF 理事长高文和阿里巴巴云智能事业群战略与合作部资深总监、达摩院院长助理刘湘雯女士为臧根林颁奖。

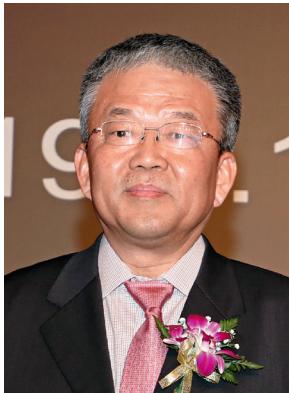
2012 年设立的“CCF 杰出教育奖”，授予在计算机教育和人才培养等方面有突出贡献的教育工作者，或在 CCF 推动中国计算机教育改革与发展方面有重要贡献的人士。2018 年，CCF 授予清华大学教授杨士强 “CCF 杰出教育奖”，以表彰他为中国高等教育发展作出的突出贡献。该奖得到了北京字节跳动科技有限公司的赞助，CCF 理事长高文和字节跳动人工智能实验室总监李航博士为获奖者颁奖。

“CCF 计算机企业家奖”于 2014 年设立，表彰在计算机和信息产业发展方面作出重大贡献的企业领导者，且企业的业绩被业内和社会广泛认可。研祥高科技控股集团陈志列先生获得 2018 “CCF 计算机企业家奖”，以表彰他对工业控制计算机产业所作出的突出贡献。CCF 理事长高文为他颁奖。

2015 年设立的“CCF 杰出工程师奖”，授予在计算机工程技术及应用领域有突出成就和重要贡献者。科大讯飞股份有限公司 AI 研究院常务副院长刘聪博士和英特尔亚太研发有限公司高级工程师吴峰光博士获得该奖，CCF 理事长高文为他们颁奖。



臧根林获CCF卓越服务奖



杨士强获CCF杰出教育奖



陈志列获CCF计算机企业家奖



刘聪、吴峰光获CCF杰出工程师奖



CCF优秀博士学位论文奖设立于2006年，每年表彰不超过10位在计算机科学与技术及相关领域有突出创新的博士学位论文的作者。为了鼓励研

究人员将高质量研究成果优先发表在中文期刊，从2018年起，“CCF优秀博士学位论文奖”将发表母语论文作为参评刚性要求。蒋炎岩（南京大学）、苗东菁（哈尔滨工业大学）、王志刚（东北大学）、易鑫（清华大学）、郑臻哲（上海交通大学）等5位在计算领域卓有建树的青年才俊获得该奖项。微软亚洲研究院一直赞助该奖项。CCF理事长高文和微软亚洲研究院常务副院长周明研究员为5位获奖者颁发了证书和奖杯。

此次颁奖会，CCF向7位新当选的CCF会士颁发了会士证书，颁发了CCF杰出员工奖、优秀专委奖、优秀会员活动中心奖、会员发展优秀奖，还为认定的五件珍贵计算机历史物件颁发了“CCF中国计算机历史记忆”认证证书。 ■



记录中国计算机发展历史 ——CCF 认定第二批“CCF 中国计算机历史记忆”

为了更好地保存中国计算机发展过程中的珍贵历史物件，中国计算机学会（CCF）于2017年开始实施“CCF中国计算机历史记忆”认定计划，认定中国研制或生产的、对中国计算机事业发展具有重要历史意义的珍贵物件，包括计算机相关的原型系统、部件、装置、书籍、软件等。

经过征集与审定，CCF历史记忆认定委员会决定认定5件珍贵的计算机历史物件为第二批“CCF中国计算机历史记忆”。在2019年1月19日举行的“2018 CCF颁奖大会”上，CCF理事长高文为这5件计算机历史物件颁发了认定证书。

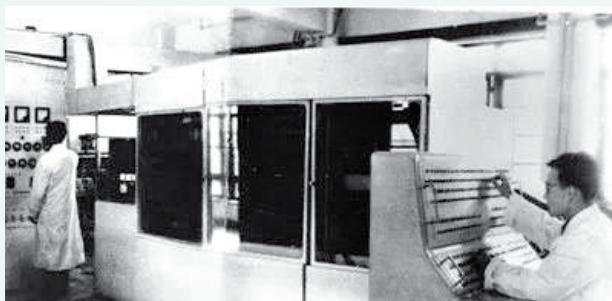
CCF 中国计算机一类历史记忆

- 西安微电子技术研究所保存的156计算机



CCF 中国计算机二类历史记忆

- 中国科学技术大学校史馆保存的“中国科学技术大学计算机专业创办历史图片和107机文档”
- 电子科技大学（原成都电讯工程学院）计算机学院保存的早期BЭСМ电子管电子计算机教材
- 电子科技大学（原成都电讯工程学院）计算机学院保存的国产441B晶体管计算机文档（讲义）
- 清华大学刘斌教授设计的64K*64K单板ISDN并行交换器



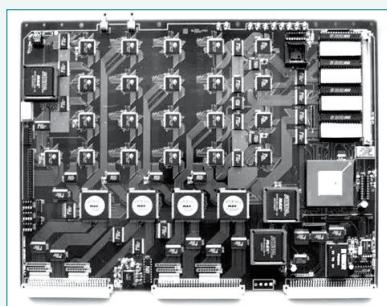
107 (KD-1) 计算机在科大安装调试



国产441B晶体管计算机文档(讲义)



早期BЭСМ电子管电子计算机教材



64K*64K单板ISDN并行交换器

CCF 将举行理事会换届选举

CCF 理事会是会员代表大会的执行机构，在会员代表大会闭会期间领导学会开展工作，理事会任期为四年。本届理事会任期将于 2020 年 1 月止，为此，2019 年 1 月 20 日举行的 CCF 第十一届常务理事会议第七次会议决定，将于 2019 年 10 月 19~20 日在苏州举行 CCF 全国会员代表大会，并将选举产生新一届 CCF 理事会。

本次常务理事会还选举产生了换届选举相关的机构和人员。

换届选举指导委员会

主席：高文 CCF 理事长，北京大学教授

委员（按姓氏拼音为序）：

卜佳俊 CCF 常务理事，浙江大学教授

杜子德 CCF 秘书长

金芝 CCF 常务理事，北京大学教授

吕建 CCF 副理事长，南京大学教授

芮勇 CCF 常务理事，联想集团 CTO 兼高级副总裁

王恩东 CCF 常务理事，浪潮集团有限公司首席科学家

臧根林 CCF 常务理事，广州科韵大数据技术有限公司创始人

周兴社 CCF 常务理事，西北工业大学教授

附：CCF 换届选举指导委员会构成和职责

一、工作职责

1. 协调各换届选举工作机构的工作。
2. 对“规章修订小组”提出的关于规章的修改意见进行确认，确认后，提交理事会或会员代表大会表决。
3. 根据《会员代表产生办法》，提名具有一定代表性的会员作为 C 类代表，交由常务理事会表决。
4. 根据学会规章和常务理事会的授权处理换届过程中除选举之外的相关事务。
5. 主持会员代表大会中除选举之外的其他议题。

二、机构成员

1. 理事长会议成员（理事长、副理事长、秘书长）及 4 名常务理事，共计 9 人，4 名常务理事中，至少一名来自工业界。
2. CCF 理事长担任主席，CCF 秘书长担任秘书。

三、任期 从 2019 年 1 月 21 日起至本次选举结束止。

换届选举规章修订组

组长：杜子德 CCF 秘书长

组员（按姓氏拼音为序）：

杜小勇 CCF 常务理事，中国人民大学教授
侯紫峰 CCF 监事，联想研究院研究员
胡事民 CCF 常务理事，清华大学教授
金 芝 CCF 常务理事，北京大学教授
马殿富 CCF 会士，北京航空航天大学教授
彭思龙 CCF 理事，中科院自动化所研究员

规章修订组的职责是修订CCF章程和与换届相关的条例；组织会员对相关事项的公开讨论并综合；将规章修订之处交由CCF换届选举指导委员会确认后，最终文本提交会员代表大会或理事会表决。

换届选举委员会主席

李晓明 CCF 会士、北京大学教授

换届选举委员会的职责是对会员代表资格进行审查并公布；对参选候选人资格进行审查并公布；主持选举，选举结束后签发证书。

会员代表产生工作组组长

彭思龙 CCF 会员与分部工作委员会主任、中科院自动化所研究员

会员代表产生工作组的职责是根据《CCF会员代表产生办法》及会员分布划分选区，任命选区负责人；将各选区产生的会员代表提交选举委员会进行资格审查。该工作组对CCF换届选举委员会负责。

CCF 常务理事会最后一轮执行委员产生

为提高决策效率和决策质量，常务理事会设执行委员会，对学会重要议题在常务理事会召开前先行进行深入研究、提出方案交由常务理事会表决。执行委员会对常务理事会负责。执行委员会设9人，其中理事长和秘书长为当然成员，其余7人由常务理事会推选产生，任期一年。在2019年1月20日召开的CCF第十一届常务理事会第七次会议上，选举以下常务理事为本届常务理事会任期内最后一轮执行委员会委员，他们是（按姓氏拼音为序）：

陈 钟 杜子德 高 文 过敏意 胡事民 史元春 孙凝晖 唐卫清 周志华

经CCF第十一届常务理事会第七次会议通过，
“CCF生物信息学专业组”升级为
“CCF生物信息学专业委员会”

七名CCF杰出会员当选CCF会士

2018年12月3日，2018年度中国计算机学会（CCF）会士评选会在CCF总部举行。本次会议由CCF会士工作委员会主席、中国科学院院士梅宏主持。2018年度共收到会士候选人提名材料15份，CCF会士评选委员会委员仔细审阅了提名材料，经过讨论和无记名投票，7位CCF杰出会员当选CCF会士。

CCF从2008年起设立会士制度，旨在表彰在计算机领域取得卓越成就或为CCF作出突出贡献并有5年以上连续会龄的CCF会员。会士是会员在CCF的最高学术荣誉。

本次当选的7名会士（按姓氏拼音为序）：



窦勇 CCF理事、体系结构专委会主任，国防科技大学研究员

窦勇研究员主要从事计算机体系结构、高性能计算机系统结构、可重构算法加速器体系结构和面向领域的大规模并行计算技术等方面的研究。参加银河III/IV/V三个型号高性能计算机和银河高性能仿真计算机研制工作，历任设计师、主任设计师和副总设计师，是国家“千万亿次高性能计算机”创新群体主要骨干成员。他长期服务于CCF，历任CCF体系结构专委会委员、秘书长、主任，致力于体系结构专委会的发展，使得专委会学术活动日益活跃，委员人数增加，并带领所在专委会积极在CNCC上举办学术活动。

杜军平 北京邮电大学教授

杜军平教授的研究领域包括人工智能理论与技术、机器学习、跨媒体大数据智能信息处理、精准搜索等，取得了多项创新性研究成果。在IEEE TKDE、TPAMI、TSMC、TNNLS、TIM、TVT、TIFS、TCST、CVPR等国际重要刊物和学术会议上发表论文365篇。他积极参加CCF组织的多种学术活动，多次参加CCF走进高校，被评为2017年度CCF杰出演讲者。作为CCF人工智能与模式识别专委会常委和CCF大数据专家委员会常委，在专委会组织的学术交流活动中担任多种学术职务，发挥了重要作用。



段振华 西安电子科技大学教授

段振华教授长期从事计算机软件和理论学科的教学与研究工作。提出了投影时序逻辑PTL，证明了命题投影时序逻辑PPTL的可判定性；建立了一个PPTL的合理且完备的公理系统。提出了基于区间的时序逻辑PPTL的模型检测理论方法；建立了并行时序逻辑程序设计语言MSVL。他积极参与CCF相关专委会的学术活动，曾任CCF佩特里专委会副主任，作为组织委员会主席承办第29届国际佩特里大会。协助筹办CCF形式化方法专委会，多次在形式化方法、嵌入式系统、软件工程和系统软件等专委会的学术活动中作学术报告。

梁吉业 CCF理事，山西大学教授

梁吉业教授长期从事计算智能与数据挖掘技术方面的教育研究工作，在粒计算与数据挖掘方面有出色的研究。在AI、IEEE TPAMI、IEEE TKDE等权威期刊和会议上发表论文200余篇。2008年至今担任CCF理事，积极协助组织CCF学术活动，担任CNCC2016组委会副主席。2012~2017年担任CCF太原会员活动中心主任，在会员发展、学术交流、CCF走进高校、推进产学研合作等方面做了大量工作。作为CCF人工智能与模式识别专委会常委，为专委的发展作出了重要贡献。





芮 勇 CCF 常务理事，联想首席技术官兼联想研究院院长

芮勇博士是多媒体及计算机视觉领域的杰出学者，ACM/IEEE/IAPR/SPIE Fellow。首次提出多媒体特征及人眼视觉感知相似度模型理论，取得图像与文本自动排版、图像搜索多角度特征重排序、基于多模态稀疏编码的图像搜索点击预测等多项开创性科研成果；获得了 ACM SIGMM 2018 技术成就奖、IEEE Computer Society 2016 技术成就奖，以及 ACM TOMM 2017、IEEE SMC 2017、ACM Multimedia 2007 等会议的最佳论文奖。他积极领导、参与 CCF 各项工作，定期在大数据等多个专委会作主题报告，积极参与 CCF 专委工委宏观布局和专委改革发展，并带领所在企业对 ADL、TF 和 YOCSEF 等活动给予了大力支持。

张 健 中国科学院软件研究所研究员

张健研究员主要研究兴趣包括自动推理、约束求解、软件测试与分析。担任《计算机学报》、JCST、Frontiers of CS、IEEE Trans. on Reliability、《中国科学》、《计算机科学与探索》等期刊的编委。国家 973 计划项目“安全攸关软件系统的构造与质量保障方法研究”首席科学家。获得中创软件人才奖、国家杰出青年科学基金。他担任 CCF 学术工委执行委员、公共政策委员会执行委员、专委工委委员；积极参与“CCF 推荐国际学术会议 / 刊物目录”的审查与修订，代表 CCF 参加专委会会议；作为 CCF 形式化方法专委会和软件工程专委会委员，承担专委会学术会议和学科发展报告撰写等工作。



朱文武 清华大学教授

朱文武教授主要从事三元空间大数据计算、视频大数据计算、社会化多媒体计算、未来多媒体通信与网络等研究工作。欧洲科学院院士、AAAS/IEEE/SPIE Fellow。在多媒体网络发展方面做出了开拓性工作，先后获得 2001 年 IEEE T-CSVT、2004 年 IEEE 通信学会多媒体通信专业委员会的最佳论文奖。在社会感知的多媒体内容分发方面做出了国际领先成果，获得 2012 ACM Multimedia（唯一）最佳论文奖。获 2018 年国家自然科学二等奖（排名第一）。在 CCCF 上组织“社会计算”专题文章。多次组织 ADL，多次在 CNCC 上组织专题论坛。2017 年被评选为 CCF 杰出演讲者。



CCF 走进高校

序号	演讲人	时 间	高 校	演讲题目
687	侯宇涛	2018.12.13	电子科技大学	深度学习入门——使用开源免费软件 DIGITS 实现手写体数字图片分类
688	陈文光 韩光洁 吴国斌	2018.12.21	大连大学	如何进行科研选题 高质量 SCI 论文撰写方法及 ESI 引用交流经验 交通大数据助力大数据学科科研发展
689	段 磊 邬向前	2018.12.22	重庆工程职业技术学院	面向相似疾病搜索的跨疾病信息网络学习 面向眼底病筛查的视网膜图像分析
690	沈 立 计卫星	2018.12.17	河北科技大学	异构多核体系结构的能效优化技术 面向 GPU 的稀疏矩阵向量乘性能优化研究

CCF 表彰 2018 年度优秀会员活动中心

无锡 广州 宁波 上海 苏州五城市分部获奖

为了更好地为会员提供本地化服务，加强会员之间的交流，提升会员服务水平，CCF于2012年开始创建以城市为单位的会员活动中心，目前数量已发展到30个。2018年，会员活动中心在会员服务和发展会员上做出了重大努力。

根据《CCF会员活动中心条例》及《2018年度CCF分部评估办法》，CCF对29个成立超过一年的会员活动中心进行了年终评估。除总部评估外，本次新增了会员满意度调查和分部互评。依据评估结果，因在会员发展、活动开展和服务会员方面成绩显著，**无锡、广州、宁波、上海、苏州**五个城市会员活动中心被评为2018年度CCF优秀会员活动中心。鉴于CCF**杭州**对CNCC2018的突出贡献，特授予其“CNCC2018特别贡献奖”。

在2018CCF颁奖大会上，CCF正副理事长为以上会员活动中心颁发了获奖证书。



获奖分部代表左起：漆锋滨、臧根林、徐建昌、吴帆、王涛、范菁

CCF 表彰 “会员发展优秀奖”

300 余位会员推荐 1500 名新会员

2018年，CCF会员数突破55000人，其中近1500名新会员来自300余位CCF会员的推荐。为表彰先进，激励会员发展，CCF评选出10位在2018年度为壮大CCF会员队伍作出突出贡献的个人，授予他们2018年度“CCF会员发展优秀奖”（名单下附）。在2018CCF颁奖大会上，CCF理事长高文，副理事长吕建、孙凝晖、王巨宏为他们颁奖。CCF会员推荐会员活动将记录每一位会员为CCF会员发展作出的贡献。多年来，已有3000余名老会员累计推荐了上万名会员加入CCF。

附：获奖者名单

王 涛 苏州蓝甲虫机器人科技有限公司
鹿泽光 中科国鼎数据科学研究院
梁启冰 日照市计算机学会 / 日照一中
金 海 华中科技大学
彭 舰 四川大学

卢惠林 无锡商业职业技术学院
郭凤广 山东省邹平市第一中学
杨晓晖 河北大学
黄建新 工信部电子第五研究所联睿公司
丁 炎 苏州倍爱斯信息科技有限公司

98位CCF专业会员晋升为高级会员

2019年1月4日，98位在各自领域做出一定成就或在CCF服务方面有突出表现的CCF专业会员晋升为高级会员。

CCF会员部共收到符合申请和推荐要求的高级会员候选人申请材料133份，CCF高级会员资格审查委员会对这些材料逐一审核讨论后，最终评选出98位高级会员。每年仅有不超过会员总数千分之五的专业会员当选为高级会员。目前CCF共有3000余位高级会员。

CCF高级会员旨在表彰那些在计算领域取得一定成就，具有至少10年专业经验、至少2年CCF连续会龄的CCF会员。CCF高级会员是申请CCF杰出会员的必要条件。

附：2018新晋CCF高级会员名单（按照姓氏拼音排序）

姓名	任职单位	姓名	任职单位	姓名	任职单位
鲍淑娣	宁波工程学院	刘 驰	北京理工大学	王中任	湖北文理学院
蔡占川	澳门科技大学	刘 锋	黑龙江工业学院	文世挺	浙江大学宁波理工学院
曹 娟	厦门大学	刘 江	美团点评	吴贺俊	中山大学
陈鸿龙	中国石油大学（华东）	刘世霞	清华大学	向永清	中科梧桐网络公司
陈 岭	浙江大学	刘婷婷	腾讯	谢 涛	UIUC
陈平华	广东工业大学	刘文懋	绿盟科技	谢雨来	华中科技大学
陈 全	上海交通大学	刘 永	南京理工大学	兴军亮	中科院自动化所
陈晓江	西北大学	刘正尧	华北计算技术研究所	徐 君	中国人民大学
陈中贵	厦门大学	罗文坚	中国科学技术大学	许倩倩	中科院计算所
程明朋	南开大学	马 天	西安科技大学	薛 云	华南师范大学
邓 成	西安电子科技大学	马望福	航空工业自控所	杨 杰	江苏省公安科技研究所
董 仕	周口师范学院	孟凡荣	中国矿业大学	杨 征	湖南天河国云公司
傅慧源	北京邮电大学	聂秀山	山东财经大学	殷昱煜	杭州电子科技大学
郭祖华	河南工学院	宁 康	华中科技大学	袁国武	云南大学
胡清华	天津大学	宁兆龙	大连理工大学	袁晓光	航天二院七〇六所
胡欣宇	山西云时代技术公司	努尔麦麦提·尤鲁瓦斯	新疆大学	袁 野	东北大学
黄大荣	重庆交通大学	潘理虎	太原科技大学	张 娇	北京邮电大学
黄婷婷	腾讯	彭 浩	浙江师范大学	张 琪	中科院计算所
黄 红	重庆邮电大学	瞿绍军	湖南师范大学	张 雷	华东师范大学
霍 珮	中科院信工所	宋 富	上海科技大学	张良杰	金蝶软件公司
蒋 浩	中科院计算所	孙大为	中国地质大学（北京）	张 胜	南京大学
金澈清	华东师范大学	孙 猛	北京大学	张 涛	哈尔滨工程大学
邝祝芳	中南林业科技大学	孙 晓	合肥工业大学	张伟哲	哈尔滨工业大学
雷 凯	北京大学深圳研究生院	汤 进	安徽大学	张 涌	中科院深圳先进院
李宝军	大连理工大学	田志宏	广州大学	赵 剑	长春大学
李 超	上海交通大学	童咏昕	北京航空航天大学	赵险峰	中科院信工所
李春国	东南大学	王 琛	华中科技大学	赵晓燕	航天二院七〇六所
李 敏	中南大学	王海波	智器云公司	周锦程	黔南民族师范学院
李小勇	国防科技大学	王嘉寅	西安交通大学	周相兵	四川旅游学院
李迎秋	大连东软信息学院	王新年	大连海事大学	周元峰	山东大学
李泽超	南京理工大学	王 鑫	长春工程学院	朱 佳	华南师范大学
梁启冰	日照市计算机学会	王 洋	山西省工信厅	祝 恩	国防科技大学
林俊聪	厦门大学			祝恒书	百度公司

26位CCF高级会员晋升为杰出会员

2019年1月4日，26位在各自领域取得突出成就或在CCF服务方面有卓越表现的CCF高级会员晋升为杰出会员。

CCF会员部共收到符合申请和推荐要求的杰出会员候选人申请材料41份，CCF杰出会员资格审查委员会对这些材料逐一审核讨论后，最终评选出26位杰出会员。目前CCF仅有200余位杰出会员。

CCF杰出会员设于2013年，介于高级会员和会士之间，旨在表彰那些在计算领域取得重大成就，具有至少15年专业经验、至少5年CCF连续会龄的CCF高级会员。每年仅有不超过会员总数千分之三的高级会员当选为杰出会员。从2019年开始，CCF杰出会员评选工作每年一次，时间安排在9月。

附：2018新晋杰出会员名单（按照姓氏拼音排序）

姓名	任职单位	姓名	任职单位	姓名	任职单位
艾萍	河海大学	李小平	东南大学	王晓阳	复旦大学
崔勇	清华大学	李学龙	中科院西安光机所	吴迪	中山大学
高宏	哈尔滨工业大学	刘敏	中科院计算所	吴亚东	西南科技大学
管海兵	上海交通大学	刘奕群	清华大学	叶保留	南京大学
郭银章	太原科技大学	卢湖川	大连理工大学	尹义龙	山东大学
韩光洁	大连理工大学	陆品燕	上海财经大学	赵东岩	北京大学
何琨	华中科技大学	田聪	西安电子科技大学	周国栋	苏州大学
黄岚	吉林大学	王国仁	东北大学	周明	微软亚洲研究院
李航	字节跳动公司	王巨宏	腾讯公司		

CCF增设五个学生分会

为了更好地输送CCF总部的资源，加强CCF学生会员的本地化服务，CCF从2011年开始建立学生分会。2018年12月至今，CCF分别在河海大学、西北工业大学、西安电子科技大学、华中科技大学和南京理工大学成立了学生分会（名单下附），众多CCF的专家志愿者出席了成立仪式，其中华中科技大学学生分会为CCF在武汉成立的第一个学生分会。至此CCF学生分会数量已达44个。

附：新成立的5个学生分会

学生分会名称	成立时间	学生分会主席	督导主任
CCF河海大学学生分会	2018年12月8日	高建	黄倩
CCF西北工业大学学生分会	2018年12月9日	任思源	於志文
CCF西安电子科技大学学生分会	2018年12月9日	李超能	苗启广
CCF华中科技大学学生分会	2018年12月23日	张信民	王多强
CCF南京理工大学学生分会	2019年1月5日	张明月	孙晋

49 名讲者被评为 2018 年度 CCF 杰出演讲者

为鼓励更多专家为 CCF 及会员服务贡献学识和智慧，并对其学术水平和贡献给予认可，CCF 从 2013 年起实施杰出演讲者计划。凡在其专业领域有相当的水平及公认的成就，并以志愿者的身份在 CCF 组织的活动上演讲并达到 CCF 所要求的力度和数量的 CCF 会员均可申请成为 CCF 的杰出演讲者。

根据《CCF 杰出演讲者计划流程及指南》，经 CCF 杰出演讲者计划工作组专家集体评议，共有 49 名专家当选为 2018 年度 CCF 杰出演讲者。

本次评审共收到经推荐和自荐产生的 600 余位专家的演讲材料，工作组对积分 10 分以上的 111 位专家进行了评议，最终评选出 49 名杰出演讲者。获得 CCF 杰出演讲者称号表明其在专业领域有相当的水平及公认的成就，并以志愿者身份在 CCF 组织的活动中演讲并达到 CCF 所要求的力度和数量，是一项崇高的荣誉。

被评为 CCF 杰出演讲者的专家将受邀为 CCF 义务演讲，亦可本人申请演讲。

附：2018 年度 CCF 杰出演讲者名单（以姓氏拼音为序）

姓名	任职单位	姓名	任职单位
卜佳俊	浙江大学	彭绍亮	国家超算长沙中心
曹文	江苏省常州高级中学	乔宇	中科院深圳先进技术研究院
陈宝权	北京大学	舒继武	清华大学
陈道蓄	南京大学	宋新波	中山纪念中学
陈恩红	中国科学技术大学	谭晓生	北京奇虎科技有限公司
陈海波	上海交通大学	唐杰	清华大学
陈文光	清华大学	王新	复旦大学
陈熙霖	中国科学院计算技术研究所	吴飞	浙江大学
陈益强	中国科学院计算技术研究所	吴国斌	滴滴出行
陈云霁	中国科学院计算技术研究所	肖依	国防科技大学
陈振宇	南京大学	徐志伟	中国科学院大学
杜军平	北京邮电大学	杨士强	清华大学
杜小勇	中国人民大学	叶国平	安徽师范大学附属中学
冯志勇	天津大学	於志文	西北工业大学
过敏意	上海交通大学	于剑	北京交通大学
韩银和	中国科学院计算技术研究所	臧根林	广州科韵大数据技术有限公司
何万青	阿里巴巴	张大庆	北京大学
华宇	华中科技大学	章文嵩	滴滴出行
金海	华中科技大学	郑宇	京东集团
金芝	北京大学	周傲英	华东师范大学
李兵	武汉大学	周明	微软亚洲研究院
李建	浙江省杭州第二中学	朱军	清华大学
李建中	哈尔滨工业大学	朱全民	长沙市雅礼中学
李曙	南京外国语学校	祝烈煌	北京理工大学
李晓明	北京大学		

读编往来

第1期卷首语《致读者》

感谢李国杰院士在学术、科研以及做人等方面给科研人员做了很好的表率，让我们体会到一个院士宽广的胸怀和思考的高度。多年来，我认真阅读了每一期《主编评语》，收获很大。《主编评语》到《卷

首语》的变革，将是一个非常好的开始。听取不同层次、不同领域的观点和心声，会让CCCF更有吸引力，更好地营造百花齐放、百家争鸣的局面。

The CS David 专栏《AI·未来》

◆ 李开复博士是一位集学术界、工业界、投资界以及战略界的大咖，《AI·未来》这本书本身就是一本AI预测AI的书籍，值得深入阅读。对科研工作者而言，需要做的不仅仅是如何将经典的AI算法应用到各个领域，还应该从科学和发展的角度，创造和发现更多兼顾可解释性和效率的新的AI算

法、操作系统以及芯片，更好地推动社会变得更智能和更美好。

◆ 中国AI的发展，必须重视日本第五代计算机计划的前车之鉴。但是，如何构建经济和社会的强制力，以便使人们更少地在意财富和权力？这个问题值得我们认真反思。

学会论坛《专委发展的历史性进步》

学会经营过程中也要谨慎对待过度商业化。CCF的发展总体是好的，能做到学术和商业化兼顾，但也有不足的地方。举个例子，如本文中的一句话“把会议就当会议了”。我们固然不能只把会议当会议，那样会缺钱，会议以后开不下去，但也一定要注重会议引入的商业模式中的用户体验。前几日出门开会，发现有一位参会者背了印有“CNCC2018”字样的背包，倍感亲切，上前搭话，他戏称“你也是花了1600元

买了个包啊”——这话有说笑的成分，然而近年来的CNCC参会收获没有以前大了也是事实。原因既不是CNCC邀请的学者差了，也不是会议的影响力降低了，而是因为近年来CNCC规模太大，导致会议的过程对与会者不够友好，影响了参会体验。

当然，做学术、做科研、做学会都不能少了钱，也不能单单指望拨款或者赞助，各级组织自身的造血功能一定要有，但要把握好“度”的问题。

各抒己见

《以“作品文化”取代“帽子文化”》

◆ “帽子”是对人才考查后的衡量结果，“帽子”本身并无问题，问题出在如何衡量的“评价文化”以及如何使用的“制度文化”。“帽子文化”盛行的原因，一方面是大多“帽子”是经过权威机构组织

专家评审而产生的；另一方面是对“作品”的评价同样存在权威性和专业性问题，操作层面上难度大，而目前大多单位并不具备对人才“作品”进行评价的资源和条件。无论如何，以“帽”取人的文化必将摒弃，专家学者自身的“作品”意识应首先树立，

相应的“评价文化”“制度文化”应该先行。

◆ 本文中提到的作品，是指具有创作性且以某种形式表现的成品。但对作品的评价机制可能不太好建立，任何一个评价体系或评价标准都难免会有弊端。我们破“五唯”，实际上破的是“唯”。至于论文也好，“帽子”也罢，本身并无过错。如果用一种成品去取代另一种成品，实际上并没有什么意义。当然，作品文化的提法是值得鼓励的。以作品说话，倒是有点像现在的预印本形式（如 arXiv），大家都把作品发上去，然后“是非审之于心，毁誉听之于人”。

《科研评价：破“五唯”，立什么？》

◆ YOCSEF能选取到当下最有争议的热点，高效组织论坛，并形成论坛纪要，这点值得点赞。虽然大家都觉得“五唯”不好，但是短时间之内还不大可能出台一个可以让大部分人接受的新的评价标准。打

破现有模式进行改革是需要很大的勇气和能力的，但是这个问题既然提出来了，我相信未来肯定会有解的。论文、职称、学历、奖项、“帽子”从不同维度衡量人的能力，最主要的问题是管理部门如何制定好的政策，合理利用这些激励人的方式，既做到普惠大部分的科研人员，又能发现和培养杰出拔尖人才，让不同层次的人感受到人文关怀，增加一些认可大家能力和价值的方式，适当引入一些市场机制，让大家各得其所，才能营造出和谐和健康的科研环境。

◆ 不一定全面破掉“五唯”，应该破“五唯”中不合理的部分，同时立新规或建立局部的补充条例。科研评价的改革是一个涉及多方面的庞大体系改革，在实行破、立的同时，要及时总结经验，不断完善评价体系，使我国的科研教育评价体系更加科学合理地运行。

建言献策

《新型存储和内存计算的回顾与展望》

这篇专题导言让我对一个不熟悉的领域快速建立了结构化认知。建议这类综述型文章有一个整体的结构图或框架图，更直观地展示所介绍方向的整体情况。建议为专题增加一个小知识或小词条模块，内容可以是相关主题的一两个关键词条的解释，相关主题的顶级科学家或研究团队的介绍，也可以是相关主题的标志性事件的介绍。

《大数据共享及交易中的机遇和挑战》

数据共享交易的前提是需要建立和完善基于隐私保护和数据安全的相应鉴定和评价组织，以期界定数据的合法性和安全性，这恰恰是困扰数据所有者的问题所在，因为其不知道数据共享交易或应用会有哪些、哪类未知风险。建议从政府层面对数据共享立法并归口相关鉴定机构，由政府出台相关数据共享交易的激励政策，最终实现市场化运营，这可能是促使未来数据共享交易的落地之举。

《电脑前传(2)：计算》

由于是系列长文刊登在不同期数的期刊上，建议在开头有个前文的简短回顾，保持连贯性，方便

读者阅读。

编辑部回复：这个建议很好。《电脑前传》后续系列文章会加一个对以前文章的简短回顾。

《大数据交易市场构建》

◆ 希望可以给出针对某一或某几个平台的分析，结合具体案例更好地诠释市场构建方法。
◆ 期望作者能够在合理的、不敏感的技术范围内，发表数据采集系统的研究报告。

2018年第11期群智协同计算专题

目前对于群智协同计算的研究大多是集中于任务的达成，大部分相关研究例如任务分配、个体的贡献、激励模式等都是为了更好地达成任务。但是在其他场景中，比如教育的群智协同中，关注点除了任务达成之外，还关注个体的成长与收获，目前与之相关的研究不多，希望能刊登此类文章，扩展群智协同的应用场景。

（本次参与评刊的有：陈盈、李挺、廖勇、刘宇擎、罗曜儒、时成阁、万江平、王波、王诗兵、易小琳、张福生、郑巍、周果）

CCF 会员活动中心动态（2018年）

CCF 大连 12月21日，CCF大连举办了2018年度学术年会。CCF副秘书长陈文光、会员与分部工委委员吴国斌、CCF计算机应用专委会常委韩光洁以及CCF大连的委员和会员近200人参加了本次年会。

12月7日，CCF大连在大连海洋大学召开了“水下机器人”学术报告会。会议邀请了大连理工大学教授、国家优秀青年基金获得者赵云鹏，大连海事大学教授、智能海洋机器人研究中心负责人王宁，大连理工大学副教授、中国科协“青年人才托举工程”、国家“香江学者”计划项目获得者杨鑫分别作了报告。

12月4日，CCF大连在大连理工大学举办“网络安全人才培养”论坛。本次活动执行主席为CCF大连副主席、大连市互联网协会副会长朱晓涛，CCF大连执行委员、大连理工大学计算机学院副院长葛宏伟。

12月1日，以“创新驱动与高质量发展”为主题的2018年大连市科学技术协会年会在世界博览广场举行。中国工程院院士郭东明，中国科学院院士邱大洪、何国钟、张东辉出席了年会开幕式。中国工程院院士、CCF理事长高文和中国科学院微电子研究所所长叶甜春受邀分别作了题为“人工智能发展现状和趋势”和“中国集成电路制造产业创新发展情况”的主旨报告。

11月29日，CCF大连在东软河口国际软件园举办“产学研合作交流会”。本次活动执行主席为CCF大连执行委员、东软集团开发中心主任南丽岚，CCF大连副主席、大连市互联网协会副会长朱晓涛。

CCF 无锡 12月21日，CCF无锡分部秘书长卢惠林一行六人赴软通动力中国区总部考察交流，受到公司技术总监郑佳和顾丽杰等的热情接待。

12月14日，由无锡市经济和信息化委员会指导，中国计算机学会(CCF)、无锡市信息化协会主办，无锡市信息安全产业联盟、CCF无锡承办的“物联网信息安全”论坛在无锡市恒华科技园举办。无锡市经信委信息安全处处长孙迎强，山水城管委员会招商局副局长张红梅、科长许元，CCF无锡分部委员周浩杰、执委杨丽参加了此次论坛。

12月7日，由中国计算机学会(CCF)主办，CCF无锡和无锡市计算机学会承办的“第七届太湖论坛——自主可控大数据与人工智能大会”在无锡举办。

CCF 合肥 12月28日，CCF合肥走进联宝(合肥)电子科技有限公司，深度了解笔记本电脑设计、制造的全过程，亲身体验国家级智能制造示范基地的现代化生产过程。

CCF 成都 11月24日，CCF成都举办全省高校计算机(软件)学院院长论坛。宜宾学院党委书记蔡乐才作专题报告，院长王玲致欢迎辞，成都信息工程大学党委书记周激流致开幕辞。来自电子科技大学、四川大学、西南交通大学等26所高校的计算机(软件)学院领导及专家70余人出席，四川华迪信息技术有限公司等五家企业代表应邀列席论坛。论坛由宜宾学院计算机学院院长李忠主持。



▲ YOCSEF
● CCF会员活动中心
▼ CCF学生分会

只有结成群体 才好发展专业

加入CCF/CCF会员资格延续

专业会员/高级会员/杰出会员/会士:200元/年(一次可交纳5年)

学生会员:50元/年

欢迎 微信支付

其他缴费方式

在线缴费 www.ccf.org.cn

银行转账

开户行: 北京银行北京大学支行

户 名: 中国计算机学会

账 号: 0109 0519 5001 2010 9702 028

