CCF 推荐 A 类国际学术会议介绍

SIGKDD二十周年庆典

唐 杰[1] 东昱晓[2] 蒋 朦[1] 方展鹏[1]等
[1]清华大学
[2]美国圣母大学

关键词：数据科学推动社会进步 数据挖掘

概况

国际知识发现与数据挖掘大会 (ACM SIGKDD Conference on Knowledge Discovery and Data Mining, SIGKDD[1]) 是数据挖掘领域的顶级国际会议。会议内容涵盖数据挖掘的基础理论、算法和实际应用。SIGKDD 发展的历史可以追溯到 1989 年开始组织的一系列关于知识发现及数据挖掘的研讨会。自 1995 年以来，SIGKDD 以大会的形式连续举办了 20 届，由于其学科交叉性和广泛应用性，影响力越来越大，论文的投稿量和参会人数逐年增加。

今年 8 月 24~27 日，第 20 届 SIGKDD [2] 在美国纽约召开。此次大会包括 1 天专题报告（tutorials 和 workshops）和 3 天主会，共吸引了 2320 人参加，几乎是 SIGKDD 2013 参会人数的两倍。

研究论文和热点

本届大会共收到 1036 篇研究性论文和 197 篇工业和政府应用性论文，均高于 SIGKDD 2013 相应投稿数量的 40%。大会最终录用 151 篇研究性论文（录用率约 14.6%）和 44 篇工业和政府应用性论文（录用率约 22%），其中包括中国大陆学者作为第一作者发表的 13 篇研究性论文。图 1 给出了自 2001 年以来 SIGKDD 每年接收的研究性论文投稿数和最终录用论文数的对比。

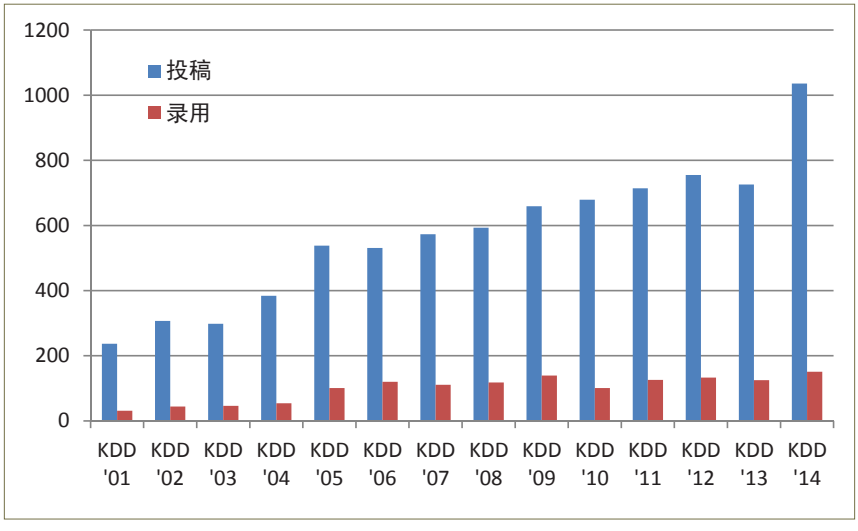

图1 历届论文投稿和录用情况

SIGKDD 有限的录用数使得对论文质量提出严格要求的同时，保证每一篇入选论文有充分的展示机会：包括 15 分钟的口头报告时间、3 分钟的问答环节以及晚间长达 4 小时的展板展示环节。论文的作者们体现出了极高的敬业精神，不仅口头报告重点突出、生动活泼，在展板展示的 4 个小时时间里也都“一站到底”；尽管口干舌燥，但依旧耐心而兴奋地介绍自己的工作成果。来自世界各地的专家

学者汇聚在一起，讨论数据挖掘领域的研究趋势，交流创新性的想法以及突破性的进展。这就是本届SIGKDD吸引2000余人的魅力所在。

在研究热点方面，社交网络、机器学习和富数据挖掘方面的投稿最多。以社交网络为例，大会主会共设有6个专题分会场，报告讨论其最新进展。表1列出了主要研究方向的投稿占整个大会投稿数量的比例以及录用率。“大数据”和“有监督学习”是本次大会论文录用率最高的两个研究方向，而“图挖掘”和“富数据挖掘”方向的论文录用率偏低。

表1 SIGKDD 2014各研究方向投稿比例和录用情况

研究方向	投稿比例	录用率
社交网络	12.84%	14%
无监督学习	8.98%	15%
富数据挖掘	8.30%	13%
图挖掘	7.63%	10%
有监督学习	7.43%	17%
大数据	6.27%	23%
数据挖掘应用	6.18%	15%
推荐系统	6.08%	14%

本届大会的主题为“数据科学推动社会进步”(Data Science for Social Good)，旨在呼吁和推动数据科学家投身和致力于解决实际社会问题。联合国官方媒体全球脉搏(United Nations Global Pulse)参加了本届SIGKDD，并从200余篇文章中选出5篇具有代表性的文章进行报道。该报道认为，这5篇文章在全球可持续发展及人道主义关怀大背景下具有现实世界的可应用性。

第一篇文章来自清华大学和美国圣母大学团队的“Inferring User Demographics and Social Strategies in Mobile Social Networks”。该文章发现了人类社交策略随着人生不同阶段社交需求变化的过程。例如，年轻人更趋向于不断扩展朋友圈，并且积极与异性保持联系；随着年龄增长，当人至中年或老年，则更多趋于保持小数量的但更亲密的同性朋友圈。基于此发现，该团队根据手机通话和短信模式准确地预测出了使用者的性别和年龄。

第二篇文章是来自IIT-CNR(The Institute of Informatics and Telematics of CNR)的“EARS (Earthquake Alert and Report System): a Real Time Decision Support System for Earthquake Crisis Management”。该团队设计了一个基于实时社交媒体数据的地震预警和损害评估系统，该系统可以实时报告地震中人口和基础设施层面的信息，并提供决策支持。

第三篇文章为东京大学的“Prediction of Human Emergence Behavior and their Mobility following Large-scale Disaster”。该文章使用日本东部大地震和福岛核泄漏事件后，160万日本人的GPS轨迹数据来分析和预测人类在紧急事件和灾难发生后的群体移动行为。

第四篇文章是来自Enigma, IBM和GiveDirectly的“Targeting Direct Cash Transfers to the Extremely Poor”。该文章通过分析肯尼亚贫困村庄的卫星遥感数据来鉴别极度贫穷家庭，以此为根据向他们提供无条件的人道主义关怀和资金支持。

最后一篇文章是来自微软亚洲研究院和上海交通大学团队的“Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City”。该论文基于北京32000辆出租车的GPS轨迹数据分析实时路况，预测不同时间和地区内该市的汽油消耗和污染排放情况。

除此之外，本次大会设置了多个分会场报告，展示了数据科学如何解决其他社会性问题，例如公共政策、医疗、智慧城市、物联网、环境、教育、交通和劳动力市场等。

审稿总结及建议

本次大会程序委员会主席卓尔·莱斯科夫(Jure Leskovec)教授总结了本届SIGKDD的论文审稿过程，并向所有参会者提出了很有启发性的建议。

作为科研工作者，自己的工作要想被SIGKDD认可，须从三方面入手。

1. 团队构成多元：团队成员不仅要有学术界人士，也要让工业界，甚至政府的研究人员参与进来。

2. 作者中至少有一名资深专家：如果有一名数据挖掘领域的资深专家在论文写作中作指导，那么论文质量会更容易达到SIGKDD的标准。

3. 不要提交超过5篇论文：当提交论文数量小

于5篇时，入稿率曲线较为平滑；但当提交论文数量超过5篇时，入稿率会急剧下滑。考虑到人的精力和时间有限，提交论文篇数过多会对文章质量有一定影响。

作为审稿人，想要共同保证SIGKDD的论文质量，要做到如下三点。

1. 尽量别给“weak reject”或是“weak accept”这种模棱两可的分数，因为中立的分数往往会给评审结果带来很大偏差。

2. 努力去写更长、更明确的审稿意见：这会与论文所得到的最终评审结果更吻合。人们往往误认为这样的审稿意见多来自年轻的审稿人，因为资深学者更忙碌，在审稿过程中更倾向于只提供积极或消极的态度。但事实上，恰恰是年长者、资深学者才会给出长而明确的审稿意见，反倒是年轻的审稿人难于开口表达看法。

3. 尽早提交审稿意见：本次大会提交审稿意见截止时，程序委员会只收到了半数的审稿意见，而在截止日结束后提交的审稿意见，无论在结果统一性还是意见质量上,都无法与按时提交的意见相比。

主题报告

本届SIGKDD邀请了5位专家作大会主题报告。

艾伦人工智能研究所首席执行官奥伦·埃齐奥尼(Oren Etzioni)博士在题为“数据挖掘未来之战”(The Battle for the Future of Data Mining)的主题报告中指出，传统的基于数据驱动的挖掘方法，包括深度学习在内，潜力有限。而未来一个重要的发展方向是通过建立大型复杂的知识库，利用知识驱动的方法来进行数据挖掘。奥伦还介绍了他的团队基于知识驱动方法开发的开放问答系统，该系统能够对美国四年级自然科学考试的题目自动给出答案。

美国艺术与科学院院士、美国工程院院士、微软雷蒙德研究院院长埃里克·霍维茨(Eric Horvitz)博士在题为“数据、预测和决策在人类和社会中的作用”(Data, Predictions, and Decisions in Support of People and Society)的主题报告中，展示了他和他的团队通过数据挖掘技术改善社会的几个项目，包括通过预测交通情况来改善城市拥堵问题，通过全球的航班数据来生成各个地区的风向图，利用病人和医院的历史数据预测哪些病人在短时间内需要再入院，通过搜索记录来检测药品副作用等。

美国伊坎基因组织学和多尺度生物学研究所所长埃里克·沙特(Eric Schadt)博士作了题为“一种基于数据驱动进行疾病诊断和治疗的方法”(A Data Driven Approach to Diagnosing and Treating Disease)的主题报告，重点介绍了他们团队发现的一种基因模块网络。该模块网络和多种疾病（包括老年痴呆症）有关联，即使在安慰剂效应(placebo effect)下依然活跃。

哈佛大学经济系教授塞德希尔·穆莱纳桑(Sendhil Mullainathan)博士在题为“社会学家对机器学习的批评”(Bugbears or Legitimate Threats? (Social) Scientists' Criticisms of Machine Learning)的主题报告中，重点强调了数据中关联关系和因果关系的问题，以及通过预测方式来检测理论的问题，同时还介绍了他们团队通过给简历随机附上黑人和白人的名字来研究美国大学毕业生失业率的工作。

彭博资讯公司董事长兼首席执行官丹·多克托洛夫(Dan Doctoroff)先生在主题报告中，介绍了彭博资讯公司使用环境、社会和政府方面的数据帮助企业客户进行投资决策的案例，强调未来如果将覆盖一个事物不同方面的各种数据整合起来使用，会对整个经济市场和人类社会产生巨大的影响。

获奖论文

SIGKDD 2014的最佳研究论文奖授予卡耐基梅隆大学及谷歌的亚伦·李(Aaron Q. Li)、阿姆鲁·艾哈迈德(Amr Ahmed)等人的论文“Reducing the Sampling Complexity of Topic Models”。该论文研究如何降低主题模型求解过程中的采样复杂度，其算法的时间复杂度比传统方法提高了一个数量级。工业和政府应用性最佳论文授予Etsy的黛安·胡(Diane J. Hu)等人的论文“Style in the Long Tail: Discover-

ing Unique Interests with Latent Variable Models in Large Scale Social E-commerce”。最佳解决社会问题论文授予 Enigma, IBM 及 GiveDirectly 的研究者撰写的论文“Targeting Direct Cash Transfers to the Extremely Poor”。此文也是前面提到的联合国官方媒体全球脉搏选出的 5 篇文章之一。

SIGKDD 从今年开始设立 Test of Time 奖，旨在表彰过去 20 年在 SIGKDD 上发表的有重大影响力的优秀论文。本次会议有三篇论文获此殊荣。第一篇论文是加拿大西蒙菲沙大学的马丁·埃斯特尔 (Martin Ester)、德国慕尼黑路德维希马克西米连大学的汉斯 - 彼得·克里格尔 (Hans-Peter Kriegel) 等人于 1996 年发表的“A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”。该论文提出的 DBSCAN 算法对基于密度的聚类算法产生了巨大影响，已成为聚类算法中公认的重要算法之一。截至今年 8 月，其 Google Scholar 引用已超过 6000 次。第二篇获奖论文是伊利诺伊大学芝加哥分校的刘兵、新加坡国立大学的魏恩·徐 (Wynne Hsu) 等人于 1998 年发表的“Integrating Classification and Association Rule Mining”。该论文率先提出了整合关联规则和分类算法来帮助数据挖掘分类的思想，激发了一系列后继工作的发展。第三篇获奖论文是南加州大学的戴维·肯佩 (David Kempe)、康奈尔大学的乔恩·克莱因伯格 (Jon Kleinberg) 和伊娃·陶尔多什 (Eva Tardos) 于 2003 年发表的“Maximizing the Spread of Influence through a Social Network”。该论文进行的社交网络中影响力最大化问题研究具有里程碑式意义，在数学上证明了求解该类问题算法的性能指标界限，为后续研究工作提供了理论支持。

今年 SIGKDD 最佳博士论文奖第一名由卡耐基梅隆大学的冈赫·吉姆 (Gunhee Kim) 博士获得，其论文题目是“Reconstruction and Application of Collective Storylines from Web Photo Collections”，导师为邢波 (Eric Xing) 教授；第二名由斯坦福大学的阿迪蒂亚·帕拉梅瓦朗 (Aditya Parameswaran) 博士获得，论文题目是“Human-Powered Data Management”，导师为赫克托·加西亚 - 莫利纳 (Hector Garcia-Molina) 教授。

SIGKDD 创新奖是数据挖掘领域最高技术奖项，每年在 SIGKDD 上颁发。今年的创新奖由华盛顿大学的佩德罗·多明戈斯 (Pedro Domingos) 教授获得，以表彰他在流式数据分析、马尔可夫逻辑网等方面的基础性工作。

致SIGKDD 20周年纪念

今年是 SIGKDD 第 20 届年会，如果加上从 1989 年开始的研讨会，SIGKDD 已有 25 岁了，这是一个意气风发的年龄。25 岁的 SIGKDD 已然获得了全球学术界、产业界以及政府机构的广泛关注和认可，成为当今以数据科学引领技术发展的一颗耀眼明星。25 年前，来自 GTE 实验室的格雷戈里·皮埃特斯基 - 夏皮罗 (Gregory Piatetsky-Shapiro) 博士在美国底特律主持了首次 SIGKDD 研讨会，这次研讨会具有历史意义，为 SIGKDD 的茁壮成长奠定了基础。

它从以下三点出发：

1. SIGKDD’89 将数据库知识发现中的关键问题一一点明，并展开热烈讨论，包括专家系统、机器学习、数据采集、智能数据库、推理和统计图等。

2. SIGKDD’89 收到的投稿来自 12 个国家和地区，9 个国家的 39 名学者参加了本次会议。

3. SIGKDD’89 收录的 9 篇文章分为数据驱动的方法、基于知识的方法和系统应用三类。

1995 年在加拿大蒙特利尔举办的首届真正意义上的 SIGKDD，吸引了来自机器学习、统计、智能数据库、数据可视化、高性能计算和专家系统等领域的专家学者。会期两天，包括特邀报告、口头报告、展板展示以及产品演示等环节。研讨会包括数据库挖掘、贝叶斯网络、模糊集、监督学习等经典问题。自此 SIGKDD 开始书写精彩篇章，开始认知世界并让世界熟识。

20 年时间里，SIGKDD 所产生的学术价值和产业价值难以估计：

1. 它提出了大量数据方面的科学问题（例如关联数据挖掘、图数据挖掘、信息网络挖掘、社交网络挖掘、大数据挖掘等），让学者甘愿倾尽一生精力致力其中。

2. 它有大量科学成果转化为实际产出（例如客户关系挖掘、基于口碑传播的营销等），让企业从数据中认知用户，让数据服务于用户。

3. 它让很多过去难以想象的生活方式成为可能，例如基于大数据的健康医疗、智能家居等。

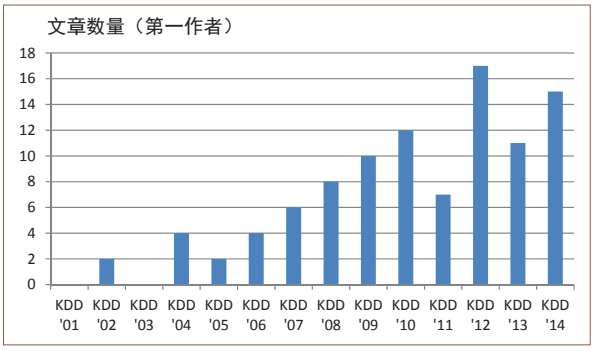

图2　历届中国大陆高校论文接收情况

值此 SIGKDD 大会 20 周年之际，我们对 SIGKDD 在中国的发展进行了分析（图 2）。自 2001 年起，中国大陆学者在 SIGKDD 大会上累计发表文章 100 余篇，其中大陆高校及科研院所共发表 98 篇，包括清华大学、北京大学、上海交通大学、复旦大学、中国科学技术大学、厦门大学及中科院等。在大会组织方面，中国大陆学者正在积极参与和投身于 SIGKDD 大会的组织工作及担任（高级）程序委员（表 2）。在大会赞助商方面，包括华为、百度、腾讯、阿里巴巴、ArnetMiner 等互联网公司和机构逐步成为 SIGKDD 大会的赞助商，在世界顶级数据挖掘盛会上宣传和推广中国信息产业和数据产业的最新研究和应用成果。整体而言，中国数据挖掘学者正在通过自己的努力和工作扩大其在 SIGKDD 大会上的影响力，并逐步得到世界同行的认可。

表2　SIGKDD大陆高校及公司参与情况

	组委会（人数）	高级程序委员会（人数）	赞助商
KDD'14	1	1	百度、华为
KDD'13	3	3	华为、ArnetMiner
KDD'12	12	4	华为、腾讯、百度、阿里巴巴、ArnetMiner
KDD'11	1	1	ArnetMiner
KDD'10	0	1	/

组委会成员寄语

在本届大会期间，本文作者对大会组织委员会成员及获奖者进行了近距离采访，邀请他们谈谈对 SIGKDD 近年发展的感受和未来展望。

SIGKDD 现任主席刘兵教授：每年的 SIGKDD 都是数据科学、数据挖掘和大数据研究的前沿阵地。人类社会正在经历一场数据革命，数据科学与大数据技术已经深入每个人的脑海，而数据处理与分析技术逐渐成为自然科学、社会科学和工程学等学科的基础技能。如何更好地利用规模庞大并以惊人速度不断增长的数据对于每一个领域和学科都显得愈发重要。作为 SIGKDD 的组委会成员，我们幸运地站在了这场数据革命的中心，站在了大数据信息时代的潮头浪尖。

作为世界首屈一指的数据科学与数据挖掘学术会议，SIGKDD 关注数据科学的基础理论性研究以及创新性模型和算法在各领域的现实应用。我们坚信学术研究和工业应用的融合与统一是未来数据科学的最佳发展方向。SIGKDD 录用的研究性论文以及工业和政府应用性论文让与会者体会到来自学术界和工业界的前沿进展，学术界和工业界科研人员可以相互学习，共同推动数据科学的发展并解决社会性问题。

SIGKDD 组织创始人之一格雷戈里·皮埃特斯基博士：今年是 SIGKDD 史上首次 50% 以上的参会者来自工业界，并且参会者来自 52 个国家和地区，创历史新高，中国有 86 名参会者，位列第二。

如何利用数学挖掘和知识发现技术帮助解决社会性问题在大会第一天即异常火爆。以我个人的观察，深度学习、社交网络与图挖掘、主题模型、推荐系统、劳动力分析等专题分会场吸引的听众最多。

今年 SIGKDD 的一个创新之处在于 Madness 大

会报告的创立，每篇论文的作者在其口头报告会的当天早上可以使用一张投影片和30秒钟的时间向所有参会者介绍和展示他们的论文。此举可以帮助与会者了解当天所有论文报告的主要内容，并解决参加哪一个分会场的报告会。

SIGKDD 2014是1989年我组织召开第一届SIGKDD研讨会后的第25个里程碑，我当年从未想到25年后SIGKDD会像今天这样壮大，让我们一起期待未来25年SIGKDD的成功！

SIGKDD 2014程序委员会联合主席、斯坦福大学卓尔·莱斯科夫教授和加州大学洛杉矶分校王伟(Wei Wang)教授：今年的SIGKDD从1036篇投稿中接收了151篇研究性论文，破纪录的投稿数量和极低的录取率让SIGKDD 2014成为过去20年间最大且最具竞争力的大会。

在众多学术会议中，SIGKDD的特别之处在于强调基于现实世界应用的学术研究。保持学术与应用的并重，能够更好地将基础算法、计算理论、数据库技术、图理论、机器学习、自然语言处理、统计学等众多领域的专业知识扩展和应用到各行各业，例如互联网、医疗、生物、商业等，从而进一步推动数据科学的进步和发展。

SIGKDD 2014创新奖获得者佩德罗·多明戈斯教授：我很荣幸能够获得SIGKDD 2014创新奖，感谢SIGKDD委员会、组织委员会以及所有专家学者的支持。

数据挖掘让我最兴奋之处在于它的多样性和实用性，机器学习、统计学、数据库、算法和系统的专家们全都致力于从数据中挖掘实用信息的目标。当今社会，大数据风暴正在席卷我们的现实生活，但任何单个大数据项目不会彻底改变我们的生活，真正的革命性变革将发生在对大数据技术日积月累后产生的质变中。与此同时，机器学习技术正伴随着大数据时代的到来而蓬勃发展，如果把数据比作燃料，计算能力比作引擎，那么机器学习技术则是让引擎工作的火花塞。

对于刚投身于数据挖掘和机器学习领域的年轻学者，我有几点建议：(1) 致力于基础性和开创性研究，莫沉迷于修补式的工作；(2) 数学服务于数据，不要被教材中看似复杂的数学吓倒；(3) 广泛浏览阅读多学科多领域文章，启发思路，开拓视野；(4) 化繁为简，用简单的方法解决复杂问题；(5) 培养和保持对数据科学研究的兴趣和乐趣。

到目前为止，我们仍然只触及大数据技术革命的冰山一角，数据科学家的缺乏是这一改革到来的主要瓶颈。我真诚地希望数据挖掘学者和从业者以及更多的普通人关注机器学习技术在数据科学中的应用，引领这场大数据技术革命。

SIGKDD 2013创新奖、Test of Time奖获得者乔恩·克莱因伯格教授：SIGKDD 2014透露出一个重要信息：以数据为中心的分析、计算方法正在从各种角度快速渗透到诸多领域，如科学研究、商务运营、政治政策以及我们生活中的点点滴滴。大会上很多论文都彰显了大会主题——数据科学推动社会进步，向我们展示了如何利用社会与城市生活所产生的海量数据，通过数据挖掘学者提出的新型挖掘方法，来解决人类社会所面临的各种各样的挑战。

本次会议还提醒学者去注意一系列重要的新问题和新的研究方向，例如数据挖掘影响药物医学和社会科学发展的潜力，以及复杂计算模型得出的结果与专家做出的判断相融合时的微妙变幻。

展望

随着大数据时代的到来，各行各业都在发生革命性变化，传统行业正在朝信息化和数据化方向加速转型，层出不穷的新兴数据产业破竹而出。今年的SIGKDD展现出数据挖掘在多学科、多领域、多产业上的广泛应用趋势，一方面数据挖掘与分析正在变为一个服务于其他科学的基础学科和基本技能，数据科学正在变为一门综合性交叉学科；另一方面，随着数字化世界的不断普及，人类社会正在产生天文数字才能描述的大数据，对大数据挖掘技术的需求正在加速增长。如何扩展数据挖掘算法，使其时间和空间复杂度能够应用在海量数据中，如何高效地将其适用于更大规模、更高速演化的网络和其他

中国电子学会访问CCF

2014年10月8日，中国电子学会秘书长徐晓兰一行八人来到CCF进行访问座谈。

CCF秘书长杜子德接待了中国电子学会同行，并与其进行交流。杜子德就中国电子学会关心的学会治理架构、会员发展、专委管理、国际合作等话题进行了介绍。徐晓兰对CCF透明的规章制度、创新的管理理念、优质的会员服务表示赞赏，并表示要向CCF学习。

杜子德（右四）与徐晓兰（右五）一行合影

CCF曾是中国电子学会下属的专业委员会，1985年独立成为全国一级学会。

结构化数据，是数据挖掘社区面临的巨大挑战。

SIGKDD 2015[3]将在澳大利亚悉尼举办。一个全新时代的SIGKDD正在提供一个更大、更高、更令人瞩目的舞台，期待中国数据挖掘领域学者的更大贡献和非凡成就。■

唐　杰

CCF高级会员、本刊编委、2012 CCF青年科学家奖获得者。清华大学副教授。主要研究方向为Web社会网络挖掘、语义Web。jietang@mail.tsinghua.edu.cn

东昱晓

美国圣母大学博士生。主要研究方向为数据挖掘、社交网络和计算社会学。
ydong1@nd.edu

蒋　朦

CCF学生会员。清华大学博士生。主要研究方向为数据挖掘、用户行为分析和社交网络分析。
mjiang89@gmail.com

方展鹏

清华大学硕士生。主要研究方向为数据挖掘、社交网络和机器学习。
fzp13@mails.tsinghua.edu.cn

其他作者：刘　兵

参考文献

[1] http://www.kdd.org/。

[2] http://www.kdd.org/kdd2014/。

[3] http://www.kdd.org/kdd2015/。