

ACM SIGKDD 数据挖掘及知识发现会议¹

清华大学计算机系 王建勇

1、 KDD 概况

ACM SIGKDD 国际会议（简称 KDD）是由 ACM 的数据挖掘及知识发现专委会^[1]主办的数据挖掘研究领域的顶级年会。它为来自学术界、企业界和政府部门的研究人员和数据挖掘从业者进行学术交流和展示研究成果提供了一个理想场所，并涵盖了特邀主题演讲（keynote presentations）、论文口头报告（oral paper presentations）、论文展板展示（poster sessions）、研讨会（workshops）、短期课程（tutorials）、专题讨论会（panels）、展览（exhibits）、系统演示（demonstrations）、KDD CUP 赛事以及多个奖项的颁发等众多内容。由于 KDD 的交叉学科性和广泛应用性，其影响力越来越大，吸引了来自统计、机器学习、数据库、万维网、生物信息学、多媒体、自然语言处理、人机交互、社会网络计算、高性能计算及大数据挖掘等众多领域的专家、学者。KDD 可以追溯到从 1989 年开始组织的一系列关于知识发现及数据挖掘(KDD)的研讨会。自 1995 年以来，KDD 已经以大会的形式连续举办了 17 届，论文的投稿量和参会人数呈现出逐年增加的趋势。2011 年的 KDD 会议（即第 17 届 KDD 年会）共收到提交的研究论文（Research paper）714 篇和应用论文（Industrial and Government paper）73 篇，参会人数也达到 1070 人。下面我们将就会议的内容、历年论文投稿及接收情况以及设置的奖项情况进行综合介绍。此外，由于第 18 届 KDD 年会将于 2012 年 8 月 12 日至 16 日在北京举办，我们还将简单介绍一下 KDD'12^[4]的有关情况。

2、 会议内容

自 1995 年召开第 1 届 KDD 年会以来，KDD 的会议内容日趋丰富且变的相对稳定。其核心内容是以论文报告和展版（poster）的形式进行数据挖掘同行之间的学术交流和成果展示。KDD 录用的论文以研究论文为主、辅以一定数量的应用论文，以及少量的系统演示论文。依附于 KDD 年会的 KDD CUP 竞赛也是会议的一项重要内容。此外，会议还包括特邀主旨报告（keynote presentations）、辅导报告（tutorials）、专题讨论（panels）、研讨会（workshops）以及工业实践及展览（Industrial practice expo track）等内容。

1. 研究主题（Research Track）

每年的 KDD 年会结束后不久，来年的会议组织者会发布论文征文通知。征文通知中会列出论文的各种投稿要求，包括会议感兴趣的主题、评价标准以及格式等。从 KDD'12 官方网站的征文通知^[5]可以了解到，KDD'12 感兴趣的研究类主题主要包括关联分析（association analysis）、分类与回归分析算法（classification and regression methods）、半监督式学习（semi-supervised learning）、聚类（clustering）、因式分解（factorization）、迁移学习和多任务学习（transfer and multi-task learning）、特征选择（feature selection）、社会网络（social networks）、图数据挖掘（mining of graph data）、时空数据分析（temporal and spatial data analysis）、可扩展性（scalability）、隐私保护（privacy）、安全性（security）、可视化（visualization）、文本分析（text analysis）、万维网挖掘（Web mining）、移动数据挖掘（mining mobile data）、推荐系统（recommender systems）、生物信息学（bioinformatics）、电子商务

¹注：本文的一个缩短版本（参见以下链接：http://dbgroup.cs.tsinghua.edu.cn/wangjy/CCCF_KDD.pdf）发表于《中国计算机学会通讯》2011 年的第 12 期。

(e-commerce)、在线广告 (online advertising)、异常检测 (anomaly detection)、以及针对大数据的知识发现 (knowledge discovery from big data) 等。论文的评价标准主要包括新颖性 (novelty)、技术质量 (technical quality)、影响力 (potential impact)、论文表达的清晰度 (clarity of writing) 等指标。

会议期间, KDD 研究论文报告按照主题会被分成了若干个分会 (session), 被录用论文的作者在相应的分会做报告。以 KDD 2011^[6]为例, 该年会的分会主题包括分类 (Classification)、矩阵分解 (Matrix factorization)、图分析 (Graph analysis)、Web 用户建模 (Web user modeling)、用户建模 (User modeling)、在线数据和数据流 (Online data and streams)、文本挖掘 (Text mining)、隐私保护 (Privacy)、社会网络 (Social networks)、理论 (Theory)、频繁集 (Frequent sets)、非监督式学习 (Unsupervised learning)、图挖掘 (Graph mining)、可扩展性 (Scalability) 和可预测建模 (Predictive modeling)。

2. 应用主题 (Industrial and Government Track)

应用主题类论文的发表和作者的与会报告是 KDD 年会的重要组成部分, 也是 KDD 相对于很多其他会议的特色之一。由于数据挖掘的广泛应用性, 应用主题类论文受到数据挖掘研究人员和开发者的重视。相对于很多其他会议, KDD 应用主题类论文的征文启事和录取更为规范。从 KDD'12 官方网站的征文通知^[5]可以看出, 该年会的企业及政府应用主题征求描述针对企业和政府部门数据挖掘解决方案的论文投稿, 并特别欢迎某些在数据挖掘技术应用过程中能够促进某些实际问题的理解或提出新的挑战性的研究问题的论文。KDD 应用主题类论文涉及的应用领域主要包括电子商务、医疗、国防、公共政策、财务、工程、环境、制造业、电信、政务等。被 KDD 录用的应用主题类论文又被分为以下 3 大类:

- 对企业、政府或其他机构带来实际价值的数据挖掘系统
- 对企业、政府或其他用户 (例如科学研究或医疗行业) 带来显著价值的知识发现
- 有潜力带来价值的前沿应用和技术

3. KDD CUP 竞赛

KDD CUP 是 ACM SIGKDD^[1]组织的有关数据挖掘和知识发现领域的年度赛事。作为 KDD 年会的重要组成部分, 自 1997 年以来已经连续举办了 15 届, 目前是数据挖掘领域最有影响力的赛事。通常每年在 KDD 会议网站上会公布当年的 KDD CUP 主题及各个子任务、数据集、考核指标等。全世界的数据挖掘参赛者在规定时间内提交解决方案和结果。优胜者名单会在 KDD CUP 网站公布, 并在会议期间颁奖。纵观历年的 KDD CUP 赛事, 我们不难发现其主题的多样性。往届的 KDD CUP 任务涉及到面向利润 (升力曲线) 优化的直接营销、计算机网络入侵检测、在线零售网站点击流分析、分子生物活性和现场蛋白质预测、生物医学文档和基因角色分类、网络挖掘与用户日志分析、粒子物理学和同调蛋白质预测、互联网用户搜索查询分类、基于图像数据的肺栓塞检测、客户推荐、乳腺癌、客户关系预测、学生成绩评估、以及基于雅虎音乐数据集的音乐推荐等众多领域。在往届的 KDD CUP 竞赛中, 某些华人组成的参赛队伍也曾取得了不俗的成绩。例如, KDD CUP 史上首次包揽了全部子项目冠军的团队就来自香港科大, 其队员包括 Dou Shen (沈抖)、Rong Pan、Jiantao Sun、Junfeng Pan、Kangheng Wu、Jie Yin、Qiang Yang (杨强)。

4. 系统演示 (Exhibit and Demo Track)

KDD 会议设有一个系统演示分会场, 用于让数据挖掘研究人员或从业者以交互的方式向与会者展示他们所开发的数据挖掘软件系统 (或组件) 的设计理念、实现技巧以及功能等。

5. 工业实践展 (Industry Practice Expo Track)

工业实践展是 KDD 从 2011 年开始新增加的一部分会议内容，主要由特邀报告和专题讨论组成。其主要目的是召集一流的业界或政府部门的数据挖掘实践者和与会者共享他们的一些有关数据挖掘应用的体会和经验。

6. 专题研讨会 (Workshops)

同其它会议类似，KDD 也设有专题研讨会，其目的是就某些热门或前沿主题让数据挖掘研究人员有机会来交流新颖的研究想法。KDD'11^[6]共设了 16 个主题研讨会。

7. 专题讨论会 (panels)

KDD 专题讨论会是就数据挖掘领域的某个重要的话题邀请几个相关的知名专家阐述自己的观点，并通过与参会者的互动来对该话题开展深入的研讨。KDD'11^[6]的专题讨论主题为“来自数据挖掘竞赛的经验教训” (Lessons Learned from Contests in Data Mining)。

8. 短期课程 (tutorials)

每年的 KDD 年会都会就几个数据挖掘主题邀请这些领域的知名专家细致地讲解与该主题相关的问题、研究背景、主要的解决方案等内容。通常每个课程的时间是半天或一天。

9. 特邀主题报告 (keynote presentations)

每年的 KDD 年会都会邀请在某些数据挖掘领域做出卓越成绩的专家学者做主题报告。以 KDD'11^[6]为例，会议共邀请了 4 位特邀主题发言人，他们分别是：美国斯坦福大学工程系教授 Stephen Boyd、Google 公司研究主管 Peter Norvig、美国加州大学圣克鲁兹分校分子生物工程系教授 David Haussler 以及加州大学洛杉矶分校计算机系教授 Judea Pearl。

表 1、2003-2011 年期间 KDD 研究论文投稿及录取情况 (注：自 2007 年以来论文不再区分长、短文，表中 2007 年以后的长、短文对应的分别是长报告和短报告)

年份	投稿	长文	长文录取率	短文	短文录取率	总录取率
2003	258	34	13.2%	36	13.9%	27.1%
2004	337	40	11.9%	44	13.0%	24.9%
2005	465	40	8.6%	36	7.7%	16.3%
2006	457	50	11.0%	55	12.0%	23.0%
2007	513	92	17.9%	N/A	N/A	17.9%
2008	510	50	9.8%	45	8.8%	18.6%
2009	537	50	9.3%	55	10.2%	19.5%
2010	578	77	13.3%	24	4.1%	17.4%
2011	714	56	7.8%	70	9.8%	17.6%
综合	4369	489	11.2%	365	8.3%	19.5%

3、 历年论文投稿及接收情况

我们对 2003 年以来 KDD 的研究类论文和应用类论文的投稿、录取率等进行了统计(注：数据来自历年 KDD 会议的论文集)。发现研究类论文的投稿量呈现逐年增加的趋势，而论文总的录取率（即包括长、短文所有被录取论文的录取率）相对稳定，基本在 17%至 19%左右徘徊。具体的分析结果见表 1。

相对于研究类论文，应用类论文的投稿量少且相对稳定，其录取率相对更高，基本在 30%左右。具体统计结果见表 2。

此外，图 1 和图 2 分别对比了 KDD 研究类论文和应用类论文的总录取率和投稿量。

表 2、2003-2011 年期间 KDD 应用（Industrial and Government）论文投稿及录取情况

年份	投稿	长文	长文录取率	短文	短文录取率	总录取率
2003	40	12	30.0%	10	25.0%	55.0%
2004	47	14	30.0%	13	28.0%	58.0%
2005	73	14	19.2%	11	15.0%	34.2%
2006	74	7	9.4%	8	10.8%	20.2%
2007	60	11	18.3%	8	13.3%	31.6%
2008	83	13	15.7%	10	12.0%	27.7%
2009	122	12	9.8%	22	18%	27.8%
2010	101	11	10.9%	9	8.9%	19.8%
2011	73	26	35.6%	N/A	N/A	35.6%
综合	673	120	17.8%	91	13.5%	31.3%

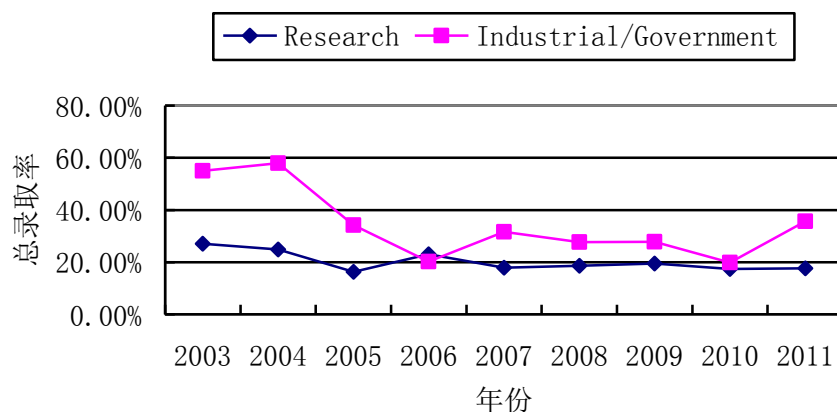


图 1、2003 年-2011 年期间 KDD 的研究论文、应用论文各自的总录取比率

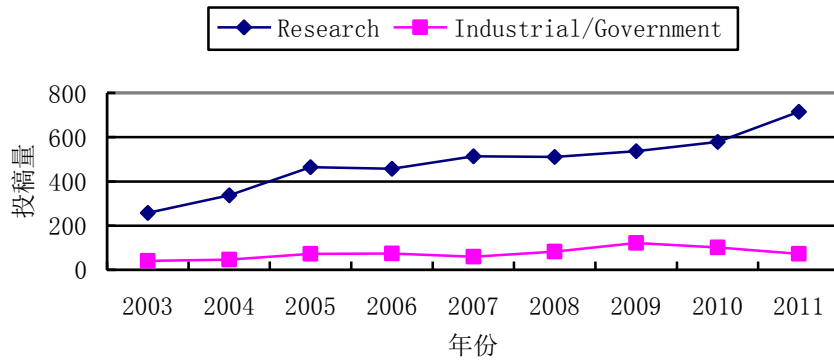


图 2、2003 年-2011 年期间 KDD 的研究论文、应用论文投稿情况

4、 设置的奖项情况

在每年的 KDD 年会上颁发的主要奖项包括 SIGKDD 创新奖 (SIGKDD Innovation Award)、SIGKDD 服务奖 (SIGKDD Service Award)、SIGKDD 最佳研究论文奖 (SIGKDD Best Research Paper Award)、SIGKDD 最佳应用论文奖 (SIGKDD Best Application Paper Award)、SIGKDD 博士论文奖 (SIGKDD Doctoral Dissertation Award) 以及 SIGKDD 学生差旅奖 (SIGKDD Student Travel Award) [3]。

1、 SIGKDD 创新奖 (SIGKDD Innovation Award)

该奖主要用于奖励对数据挖掘及知识发现领域作出重大技术贡献的研究人员, 获奖人员的研究工作通常在数据挖掘理论或商业数据挖掘系统开发上产生了持久的影响。自 2000 年以来已有 11 位数据挖掘研究人员获此殊荣, 其中来自 UIUC 的韩家炜教授位列其中。

2、 SIGKDD 服务奖 (SIGKDD Service Award)

该奖主要奖励对数据挖掘及知识发现领域作出重大服务贡献的个人或团队, 考察的因素主要包括主持学术团体、主办会议等服务性工作、数据挖掘教学、财务赞助等。自 2000 年以来已产生了 10 位获奖者, 其中包括来自美国佛蒙特大学的吴信东教授。

3、 SIGKDD 最佳研究论文奖 (SIGKDD Best Research Paper Award)

该奖主要用于奖励从每年的 KDD 年会所录用的研究论文中挑选出来的、对数据挖掘和知识发现领域有基础性的推动作用的论文。KDD 的研究论文奖包括最佳研究论文奖 (Best Research Paper Award) 和最佳学生论文奖 (Best Student Paper Award) 两类。

4、 SIGKDD 最佳应用论文奖 (SIGKDD Best Application Paper Award)。

该奖主要用于奖励从每年的 KDD 年会所录用的应用论文中挑选出来的、能较好体现在数据挖掘应用中反映出挑战性的研究问题和经验教训的论文。

5、 SIGKDD 博士论文奖 (SIGKDD Doctoral Dissertation Award)

该奖项是从 2008 年开始设立, 用于奖励在数据挖掘与知识发现领域作出出色研究工作的博士生。本科毕业于清华大学、来自美国 UIUC 的 Xiaoxin Yin 博士 (导师为韩家炜教授) 曾获得首届 SIGKDD 博士论文奖。

6、 SIGKDD 学生差旅奖 (SIGKDD Student Travel Award)

该奖项主要用于资助部分参会学生的差旅开销。

5、 关于 KDD 2012

KDD 2012 年会将于 2012 年 8 月 12 日至 16 日在北京举办,这也是 KDD 首次在亚太地区举办。中国近年来的快速发展举世瞩目。数据挖掘作为一个各个行业发展不可缺少的技术支持,在中国得到了长足发展。KDD 2012 对 KDD 以及中国的数据挖掘都是具有重大意义的里程碑。海内外数据挖掘领域的华人学者在 KDD'12 的组织工作中扮演了重要角色。例如,大会主席是香港科技大学的杨强教授,大会荣誉主席为中科院的陆汝钤院士和清华大学的张钹院士,大会指导委员会主席为中国电子工程系统研究所的李德毅院士,大会副主席为 CityGrid Media 的沈抖博士,加拿大西蒙弗雷泽大学(SFU)的裴健教授、美国罗格斯大学 (Rutgers University) 的熊辉教授和微软的 Ying Li 博士分别担任大会程序委员会联合主席、企业及政府应用分会程序委员会联合主席和工业实践展程序委员会联合主席。专题研讨会联合主席包括南京大学的周志华教授,会议会务主席(local arrangement chair)由清华大学的唐杰博士担任。相对于往届的 KDD 会议,KDD'12 的一个特色是新增加了“亚太主题分会”(Asia Pacific Track)。亚太主题分会的主席为香港大学的张伟牢教授和美国北卡大学 (UNC) 的王蔚教授。该分会将邀请亚太地区在数据挖掘领域的某些知名专家做特邀报告。与工业实践及展览分会类似,亚太主题分会不准备以论文的形式进行。此外,KDD 2012 还将举办 KDD 暑期学习班,邀请数据挖掘的知名学者就某些专题进行详细的讲解。

6、 总结

数据挖掘是一个较新的交叉学科,近年来随着海量数据在各个行业的涌现,发挥了越来越大的推动作用,受到了广泛的关注。全球的华人学者在这一研究领域扮演着举足轻重的角色。国内也孕育出了一支庞大的数据挖掘研究及开发队伍,并且在最近几年的 KDD 年会上有出色的表现。北京 KDD 2012 将为全球的科研工作者提供一个了解和学习中国数据挖掘进展的机会,也为国内的学者提供一个学习和展现的机会。这必将成为数据挖掘研究与应用发展的一个新的里程碑。最后,预祝 2012 年 ACM SIGKDD 国际数据挖掘年会取得圆满成功。

7、 致谢

十分感谢香港科技大学的杨强教授和 CityGrid Media 的沈抖博士在本文撰写过程中所给予的悉心指导和宝贵建议。

8、 参考文献

1. <http://www.kdd.org/>
2. <http://www.kdd.org/kddcup/index.php>
3. <http://www.kdd.org/awards.php>
4. <http://www.kdd.org/kdd2012/>
5. <http://www.kdd.org/kdd2012/cfp.shtml>
6. <http://www.kdd.org/kdd2011/>