数据可视分析挑战赛三年回顾

赵颖[1]，张卓[2]，袁晓如[3]
（1. 中南大学信息科学与工程学院，湖南 长沙 410075；
2. 360 企业安全集团，北京 100016；
3. 北京大学机器感知与智能教育部重点实验室，北京 100871）

摘　要：数据可视分析挑战赛作为我国可视化与可视分析领域的一项重要赛事，经过 2015~2017 三年发展，已成为推动领域竞技交流、人才培养和实践创新的重要力量。首先阐述了数据可视分析挑战赛产生的背景和意义，然后从组织形式、比赛内容和信息统计等方面对挑战赛进行了全面的回顾和分析，最后对挑战赛的未来发展进行了展望。

关键词：可视化；可视分析；学科竞赛；数据可视分析挑战赛
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.2096-109x.2018012

ChinaVis Data Challenge from 2015 to 2017

ZHAO Ying[1], ZHANG Zhuo[2], YUAN Xiaoru[3]
1. School of Information Science and Engineering, Central South University, Changsha 410075, China
2. 360 Enterprise Security Group, Beijing 100016, China
3. Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing 100871, China

Abstract: The evolution of the ChinaVis Data Challenge from its origin 2015 to 2017 was presented. Firstly, the background and organization of the Data Challenge were introduced. Then, the competition contents and statistical data of the three years were summarized. Finally, the future development of the Data Challenge was discussed.

Key words: visualization, visual analytics, discipline contest, ChinaVis Data Challenge

1　引言

竞赛是各学科领域完善其教育和科研体系的重要组成部分。对教学科研来说，竞赛具有示范和导向效应，可以丰富教学方式和引领科研方向；对人才培养来说，竞赛为参与者提供锻炼和展示自我、与同行竞技交流的平台；对学科发展来说，竞赛推动知识和技术共享，促进实践与创新融合。以网络安全和数据挖掘领域为例，网络安全领域已形成夺旗类、破解类和培养类百花齐放的国际化竞赛体系[1,2]，有效地促进了学术界与工业界的协同发展；数据挖掘领域顶级学术会议 KDD 每年举办的 KDD Cup[3]被誉为大数据分析“奥运会”，从 1997 年至今已有二十几年历史，见证和推动了数据挖掘领域的发展与繁荣。

近年来，我国在可视化与可视分析领域呈现出较为均衡且加速发展的态势。各高校和研究院的可视化与可视分析研究取得了令人瞩目的进步，尖端研究初具国际影响，基础教育日益扩大，更多单位开始关注并加入其中。工业界正加速布

收稿日期：2017-12-05；**修回日期**：2018-01-13
通信作者：赵颖，zhaoying@csu.edu.cn
基金项目：国家自然科学基金资助项目（No.61402540, No.61672538）
Foundation Item: The National Natural Science Foundation of China (No.61402540, No.61672538)

局相关领域的技术研发与行业应用，阿里和百度分别推出了组件式可视化基础开发库 DataV[4]和 ECharts[5]，360 企业安全[6]和海云数据[7]等企业也纷纷推出了面向行业的大数据可视分析解决方案。与此同时，可视化技术也逐步被广大民众熟悉与接受，2016 年“可视化推动大数据平民化”位居中国计算机学会大数据专家委员会评选的大数据发展十大趋势之首[8]。

在此背景下，我国各界有志之士积极借鉴国际赛事的经验，在国内逐步发起各类可视化与可视分析竞赛，以进一步推动领域人才培养和实践创新。效仿国际著名数据分析竞赛平台 Kaggle[9]，阿里云在国内创办了天池大数据竞赛平台[10]，并于 2016 和 2017 年分别推出了 2 个数据可视化竞赛项目“公益云图数据创新大赛”和“广东政务数据创新大赛”，2 项赛事连续成功举办，使其成为国内工业界可视化大赛的标杆。借鉴国际顶级会议 IEEE VIS Conference[11]每年举办国际可视分析挑战赛 VAST Challenge[12,13]的成功经验，国内规模最大的全国性年度可视化与可视分析技术及产业应用大会 ChinaVis[14]，于 2015 年首次设置数据可视分析挑战赛（ChinaVis Data Challenge）[15]，三年来，挑战赛飞速发展，已成为学术界主导的国内可视化顶尖赛事。

本文对数据分析挑战赛三年发展过程进行回顾与总结。首先介绍挑战赛的产生背景和组织形式，帮助读者快速了解挑战赛；然后对三年挑战赛的报名情况、评审情况、开发工具使用情况等参赛者普遍关心的问题进行解答，希望能帮助参赛者快速融入比赛节奏；同时还提供了这三年挑战赛题目与数据信息以及其他相关竞赛信息，希望能促进各类竞赛互相学习和共同进步，更广泛、更长远地推动领域人才培养和实践创新；最后对挑战赛未来发展进行展望。

2　挑战赛起源

数据可视分析挑战赛依托于每年召开的中国可视化与可视分析大会。该大会前身是由北京大学机器感知与智能教育部重点实验室主办的可视化研讨会，从 2008 年到 2013 年间共举办了 4 次。随着可视化与可视分析研究与应用在社会各界的影响不断加深，为促进中国及周边地区相关领域产、学、研、用协同发展，由我国可视化业界工作者联合发起了中国可视化与可视分析大会，分别于北京（2014 年）、天津（2015 年）、长沙（2016 年）和青岛（2017 年）成功举办了 4 届，大会每年吸引了国内外数百名可视化研究人员与领域专家汇集一堂，共同探讨在大数据与人工智能时代，可视化与可视分析领域面临的机遇与挑战。

中国可视化与可视分析大会于 2015 年首次设立数据可视分析挑战赛，至今已成功举办 3 届。作为大会的一个重要环节，挑战赛提供一系列精彩的剧本、详实的数据和具体的问题，并邀请研究人员、开发人员和爱好者使用他们最有效的可视分析技术和工具完成数据分析任务。挑战赛旨在为参赛者提供锻炼自我和竞技交流的平台，帮助参赛者评估他们的技术和工具在解决复杂问题中的有效性和新颖性，推动可视化与可视分析的专业人才培养，促进相关研究与应用的发展与进步。

3　挑战赛组织形式

挑战赛的组织工作每年与 ChinaVis 筹备同步进行，总体分为题目准备、发布宣传、公开报名、作品提交、综合评审、会场交流 6 个阶段。

题目准备阶段一般每年年初开始，工作内容包括成立挑战赛组委会、确定赞助商、讨论组织方案、准备比赛题目。360 企业安全集团是 2015 年和 2016 年挑战赛的独家赞助商，2017 年挑战赛由 360 企业安全和海云数据共同赞助。赞助商不但为挑战赛提供奖金和奖品，还在题目准备、发布宣传、综合评审和会场交流多方面提供技术支持。近三年挑战赛题目背景和数据都来自真实场景，一般先由赞助商提供样本数据和总体设想，然后组委会与赞助商技术人员一起确定题目细节，并根据题目设置对数据集进行清洗、脱敏、加工和验证性分析。

发布宣传与公开报名阶段在每年 4 到 5 月进行。挑战赛组委会首先会在 ChinaVis 官方网站上发布当年赛题、答卷和数据集，然后通过微博、微信和其他学术交流平台进行宣传。参赛者以组

队形式网上报名，全国普通高等学校和科研院所的师生和研究人员、企事业单位的开发人员和设计师、爱好者都可以报名参加挑战赛。

作品提交时间一般在 ChinaVis 召开前一个月左右，提交要求与 VAST Challenge 非常类似，每个参赛队必须同时提交答卷、视频和论文。答卷要求参赛队在完成对数据的综合分析后，用图文并茂的方式尽可能全面和准确地回答题目中预设的问题。视频要求参赛队通过影音手段解释其分析方法和流程。论文则是希望参赛队用不超过两页篇幅总结其可视分析方案的亮点。

所有参赛作品同时提交给可视分析专家和领域专家进行通信评审，评审重点关注参赛作品对预设问题解答的准确性以及可视分析方案的有效性，同时还会从新颖性、可扩展性和文档质量 3 个方面对作品进行综合评价。挑战赛委员根据专家通信评审结果进行讨论后，评选出一等奖、二等奖、三等奖和优秀奖作品，并邀请所有获奖参赛队来 ChinaVis 进行现场交流。

挑战赛在这三年的 ChinaVis 期间都安排了主会场、分会场和海报 3 个不同的交流环节。一等奖参赛队将获得在大会主会场报告的机会，同时组委会和出题方代表也将在主会场介绍本年度比赛总体情况和题目背后的故事。分会场环节包括特邀报告、部分获奖作品分享、现场讨论和颁奖仪式。海报环节则为所有获奖参赛队提供更自由和深入的交流机会。ChinaVis 结束后，组委会会择机公布当年赛题的参考答案和获奖作品，使其能够在一定程度上成为业界科研与应用研发的标准测试数据集。

4　挑战赛三年内容回顾

本节分别介绍 2015 至 2017 三年挑战赛的比赛内容，包括赛题背景、相关数据和分析需求。

4.1　2015 年挑战赛题目介绍

2015 年挑战赛题目[15]以多源异构大数据分析为主题，故事背景发生在一家为企业和政府机构提供定制化网络服务的互联网公司。该公司内部网络平台上运行着很多业务系统，这些业务系统每天都会产生很大流量，不幸的是，某几天公司某个内部系统出现了安全问题，导致该公司的部分重要数据泄露。该公司想了解那几天到底发生了什么事情，因此将那一段时间的多种网络安全日志数据提供出来，希望参赛者能以网络安全分析人员的身份，采用可视分析技术对这些数据进行分析，帮助该公司找到答案，并提供网络安全改进方案。

挑战赛在上述业务场景下设置了两道比赛题目。第一题关注公司内网正常运行模式分析，提供了一周约 700 MB 的 Tcpflow 日志数据，希望参赛者准确找到公司内部网络的客户端和服务器，总结公司内网有哪些正常网络通信模式。第二题关注安全事件分析，提供了两周共 10 GB 的 4 种网络安全日志数据（Tcpflow 日志、HTTP Header 日志、数据库行为日志、各应用系统登录日志），希望参赛者能分析出这次安全问题发生的原因、相关网络资源、演化过程以及造成的危害。

4.2　2016 年挑战赛题目介绍

2016 年挑战赛[16]设置了 2 个背景完全不同的题目。第一个题目是对 2015 年挑战赛题目的延续，该题仍以某互联网公司的内网安全为背景，提供的仍是 Tcpflow 网络流量日志数据。不同的是这次 Tcpflow 数据的信息更为丰富，记录了网络数据流在应用层、网络层和链路层 3 个层面的信息，时间跨度 2 个月，数据量 230 MB。参赛者被要求分析流量数据中隐含的正常模式和异常事件，并设计新颖的可视化方案，有效展示多层次网络行为模式。

第二题背景来自网络安全领域一个非常著名的真实事件。Hacking Team[17]是一家意大利米兰的信息技术公司，该公司向客户提供信息系统入侵、文件解密与通信监视服务。2015 年 7 月 5 日，Hacking Team 公司被黑客入侵，入侵者通过公司官方 Twitter 账号公布了该公司许多内幕信息并通告该公司的内部数据已经泄露，近 400 GB 被公开的内部电子邮件、各种相关文件和源代码引起了业界的一片哗然。挑战赛从这批数据中提取了公司内部十多年约 400 MB 的邮件数据作为比赛数据集，邀请参赛者通过分析邮件数据，解密 Hacking Team 公司的组织结构，分析该公司的发展历程及各发展阶段的业务特点。

4.3　2017 年挑战赛题目介绍

2017 年挑战赛[18]以时空数据分析为主题，设置了两道赛题，分别对应 2 个不同社会热点问题。第一题的背景是伪基站行为分析。伪基站即假基站，能够搜集以其为中心、一定半径内的手机卡信息，利用 GSM 验证漏洞伪装成真基站，冒用银行、运营商、国家机关或他人号码，强行向用户发送诈骗、色情、赌博、广告等垃圾短信，这不仅干扰公共频率资源，影响正常通信，而且严重侵害了社会秩序。手机卫士软件可以根据用户举报的垃圾短信信息，初步确定当时伪基站的近似位置，但伪基站流动性强，依据近似位置和传统数据分析方法，仍然很难准确把握伪基站的活动规律。本题提供北京市 2 个月约 753 MB 标注为伪基站发送的垃圾短信样本数据，希望参赛者鉴别不同垃圾短信类型，并分析其相关伪基站时空活动规律，为执法人员打击整治伪基站出谋划策。

第二题背景是关于某市黑网吧整治行动。随着网络宽带进入千家万户，网吧经营举步维艰，于是部分经营者不惜采用接纳未成年人、通宵营业等违规经营方式吸引顾客。某市正在开展严查黑网吧的专项执法行动，但网吧数量庞大，传统挨个排查方法不仅耗时耗力，还为黑网吧提供了规避查处的准备时间。为了提升全局掌控能力和执法效率，某市公安局将市内 3 000 多个网吧基本信息及 3 个月约 1.7 GB 上网记录数据提供出来，希望参赛者从数据中识别不同上网人群，探索上网人群的时空行为特征，检测非法上网行为与团伙上网行为，为甄别黑网吧提供线索，为犯罪预防和维护社会公共安全提供新思路。

5　挑战赛三年统计分析

本节从报名情况、评审情况和其他情况 3 个方面对 2015 至 2017 年挑战赛的相关信息进行统计与分析。

5.1　报名情况

挑战赛要求参赛者以组队方式报名，图 1 显示了这三年的报名队伍数量和队伍类型所占比例。三年来，挑战赛的影响力迅速扩大，报名队伍数量保持高速增长，2017 年相比 2015 年增加了三倍多。如果某个参赛队的成员以学生居多，则将这支参赛队当作学生队，否则当作企业工程师或爱好者队。这三年参赛队绝大多数都是学生队，其中专科生占 1%，本科生占 35%，硕士生占 42%，博士生占 4%，指导老师占 18%。参加形式多样的学科竞赛对青年学生来说有着多方面意义，因此依托学术会议举办的数据分析挑战赛，学生顺理成章地成为主力军。

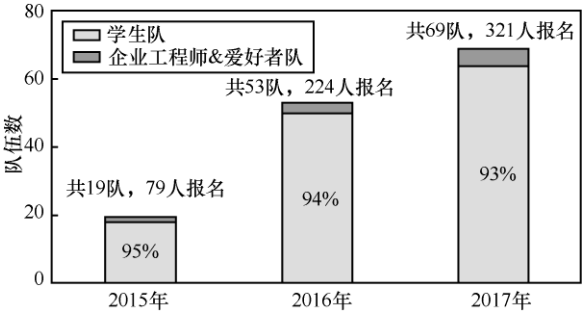


图 1　挑战赛报名队伍数量与队伍类型三年统计

企业工程师和爱好者队 2015 年有 1 支，2016 年有 3 支，2017 年达到了 9 支，虽然总体占比很少，但仍是挑战赛的亮点之一。特别是企业工程师队，他们一般会根据其行业背景非常有针对性地选择赛题，能给人们带来更深入的见解，2017 年有 2 支获奖企业工程师队的日常工作都与赛题背景有关。企业工程师和爱好者越来越多地参与进来，反映了挑战赛正在逐步引起工业界更多的关注，有利于促进学术界与工业界间多种形式的互通与交流。

5.2　评审情况

正式提交进入评审阶段的作品数量一般要比报名队伍数少，2016 和 2017 年报名队伍的提交比例大概在 60%左右，2015 年提交比例最高，19 支报名队伍中有 14 支提交作品，这与第一届比赛只有相对小范围人群关注有关。在作品提交数量和获奖率上，本文将挑战赛与 VAST Challenge 进行比较分析。图 2 和图 3 分别显示了挑战赛和 VAST Challenge 这三年的作品提交与获奖情况。挑战赛作品提交数量增长很快，而 VAST Challenge 作品提交数量呈现较大幅度波动，挑战赛的获奖率明显高于 VAST Challenge 获奖率。一方面反映出 VAST Challenge 参赛者具有更高的专业水平，另一方面也因为在创办初期的挑战赛希望通过适当提高获奖比例来吸引更多参赛者。

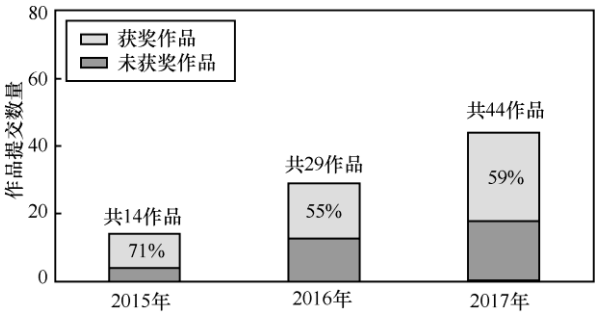


图 2　挑战赛作品提交数量与获奖比例三年统计

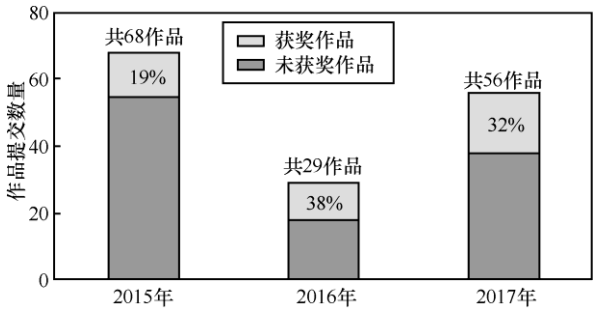


图 3　V AST Challenge 作品提交数量与获奖比例三年统计

挑战赛通信评审采取分项打分形式，每位评审专家要对所评作品从分析质量（解答题目预设问题的准确性）、可视化设计、交互设计、新颖性、可扩展性和文档质量 6 个方面做出量化评价，5 分制评分，1 分为最差，5 分为最好，这 6 个方面的权重占比分别是 30%、20%、20%、10%、10%和 10%。图 4 显示了三年所有提交作品分项打分统计情况，包括每项的平均分和方差。从平均分方面看，参赛者普遍提交了较为整齐、清晰和可读性高的作品文档与视频，因此文档质量平均分明显高于其他评分项；挑战赛这三年每个赛题都提供了参考答案，评审专家可以较客观地对分析质量打分，平均分较好的分析质量反映了参赛者都能认真地理解题目需求和努力地分析数据，并较好地解答了题目预设问题；可视映射与交互功能是可视分析系统相辅相成的 2 个最重要的组成部分，交互设计平均分明显低于可视化设计，反映了许多参赛者把更多精力投入到了呈现和解释数据方面，而忽略了交互功能带来的分析和理解作用；新颖性平均分也相对较低，反映了提出和实现有创新性的可视化与交互设计及新算法，对大部分参赛者难度较大。从方差方面看，交互设计和新颖性 2 个打分项的方差明显高于其他 4 项，结合这 2 项平均分也相对较低，说明它们是目前评价体系中对作品质量区分度最大的评分项。

5.3　其他情况

在其他情况分析中，本文重点讨论开发工具使用和数据下载情况。本文按年统计了所有提交作品用到的开发工具，表 1 给出了每年使用次数排名靠前的开发工具。D3（data driven document）[19]连续三年排名榜首实属意料之中，强大的定制能力和方便的数据驱动模式，使它成为当前可视分析领域复杂系统开发与前沿技术研究的首选工具。同为 BI（business intelligence）工具的 Excel 和 Tableau[20]出镜率都很高，其中，Excel 因微软 Office 系列工具在国内更为普及，排名一直比 Tableau 靠前。在数据处理方面，MySQL 是数据存储首选工具，Python 则成了算法处理首选。Gephi[21]的使用率高，主要因为这三年赛题都与网络分析有关。非常值得一提的是开源可视化库 Echarts[5]，它从 2015 年和 2016 年的第七名飞跃至 2017 年的第二名，这是国产可视化工具甚至国产开源软件的骄傲。

每年挑战赛数据被下载次数要远大于报名队伍数量，这间接反映了挑战赛不但为参赛者提供

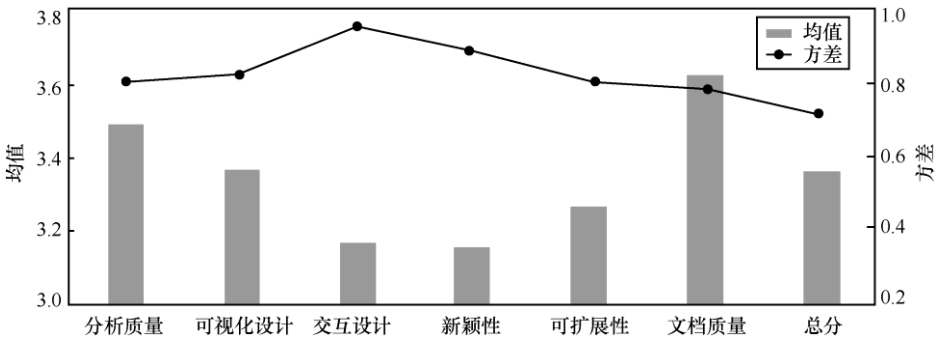


图 4　挑战赛作品评审各评分项的平均分与方差统计

表 1　挑战赛作品使用开发工具统计排名

2015 年		2016 年		2017 年	
工具	次数	工具	次数	工具	次数
D3	6	D3	21	D3	27
Excel	5	Excel	21	Echarts	26
Gephi	5	MySQL	16	Excel	23
MySQL	5	Gephi	14	MySQL	18
Tableau	4	Python	11	Tableau	13
Python	4	Tableau	9	Python	7
Echarts	2	Echarts	7	Gephi	7

锻炼和交流的机会，还为广泛的业内人士和爱好者提供了背景故事和标准测试数据集。据不完全统计，2016 年第一题数据在比赛期间就被来自 961 个不同 IP 地址的用户下载了 1 120 次，第二题数据被下载了 607 次；2017 年第一题数据在比赛期间被来自 380 个不同 IP 地址的用户下载了 556 次，第二题数据则被下载了 305 次。在比赛完成后，这些数据仍持续被下载，据不完全统计，2016 年的第一题数据在赛后 5 个月内又被累计下载了 400 多次。

6　结束语

数据可视分析挑战赛吸引了众多学生、教师、企业工程师和爱好者，并为他们提供了一个竞技、交流、学习的舞台。这三年，挑战赛在规模增长和影响力扩大等方面取得了令人鼓舞的成绩，笔者收到了参赛者和关注者大量正面积极的评价，也收到了很多非常好的建议。今后几年，挑战赛将进一步吸引学会、政府部门及社会各界有志之士以多种形式参与进来，努力扩大国内甚至国际影响力。笔者还将不断完善和丰富比赛组织形式，例如，设置应用创新类或创意设计类题目、设置单项奖励、优化赛程和评审机制等。总之，笔者仍然非常看好数据可视分析挑战赛的发展前景，也期待与社会各界一起努力，共同推动可视化与可视分析领域竞赛体系建设，促进领域的发展与进步。

致谢

感谢中国可视化与可视分析大会指导委员会对挑战赛筹备的大力支持；感谢 2015 年及 2016 年挑战赛共同主席新疆大学努尔布力、中国科学院时磊、天津大学张亚平和国防科大大学宋汉辰对挑战赛组织工作的贡献；感谢 360 企业安全的黄鑫、黄伟、李秋生，海云数据的李龙辉和华德禹，西南电子电信技术研究所的王劲松为挑战赛提供技术支持；感谢所有通信评审专家在百忙之中为参赛作品给出中肯细致的评价；感谢所有参赛者和参会者的热情参与。

参考文献：

[1] 诸葛建伟. 我国网络空间安全技能竞赛现状漫谈[J]. 中国计算机学会通讯, 2016, 12:(6):49-50.
ZHUGE J W. Discussion on the present situation of Chinese cyberspace competition[J]. Communications of the CCF, 2016, 12(6): 49-50.

[2] 余翔湛, 张宏莉, 于海宁, 等. 网络空间安全竞赛及人才管理[J]. 中国工程科学, 2016, 18(6):49-52.
YU X Z, ZHANG H L, YU H N, et al. Cyberspace security competition and talent management[J]. Engineering Sciences, 2016, 18(6): 49-52.

[3] KDD Cup 2017 homepage[EB/OL]. http://www.kdd.org/kdd2017/calls/view/kdd-cup-2017-call-for-proposals.

[4] DataV homepage[EB/OL]. https://data.aliyun.com/visual/datav.html.

[5] ECharts homepage[EB/OL]. http://echarts.baidu.com.

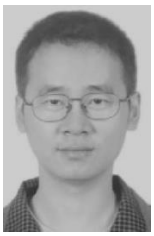
[6] 360 企业安全集团主页[EB/OL]. http://www.360.net.
360 Enterprise Security Group Homepage[EB/OL]. http://www.360.net

[7] 海云数据主页[EB/OL]. http://www.hiynn.com/hy-zh.
HYDATA Homepage[EB/OL]. http://www.hiynn.com/hy-zh.

[8] 潘柱廷, 程学旗, 袁晓如, 等. CCF 大专委 2016 年大数据发展趋势预测—解读和行动建议[J]. 大数据, 2016, 2(1):105-113.
PAN Z T, CHENG X Q, YUAN X R, et al. Developing trend forecasting of big data in 2016 from CCF TFBD: interpretation and proposals[J]. Big Data Research, 2016, 2(1):105-113.

[9] Kaggle homepage[EB/OL]. https://www.kaggle.com.
[10] 阿里天池主页[EB/OL]. https://tianchi.aliyun.com.
Tianchi homepage[EB/OL]. https://tianchi.aliyun.com.
[11] IEEE VIS conference homepage [EB/OL]. http://ieeevis.org
[12] VAST challenge homepage[EB/OL]. http://www.vacommunity.org.
[13] COOK K, GRINSTEIN G, WHITING M. The VAST challenge: history, scope, and outcomes: An introduction to the Special Issue[J]. Information Visualization, 2014, 13(4):301-312.
[14] 中国可视化与可视分析大会主页[EB/OL]. http://chinavis.org.
ChinaVis conference homepage[EB/OL]. http://chinavis.org.
[15] ChinaVis data challenge 2015 Homepage [EB/OL]. http://chinavis.org/2015/challenge.html.
[16] ChinaVis data challenge 2016 Homepage [EB/OL]. http://chinavis.org/2016/challenge.html.
[17] Hacking team Wiki homepage[EB/OL]. https://en.wikipedia.org/wiki/Hacking_Team.
[18] ChinaVis data challenge 2017 homepage[EB/OL]. http://chinavis.org/2017/challenge.html.
[19] D3 homepage[EB/OL]. https://d3js.org.
[20] Tableau homepage[EB/OL]. https://www.tableau.com.
[21] Gephi homepage[EB/OL]. http://gephi.org.

[作者简介]



赵颖（1980-），男，湖南益阳人，博士，中南大学副教授，主要研究方向为信息可视化与可视分析。

张卓（1988-），男，山东青岛人，360企业安全集团高级工程师，主要研究方向为高级威胁检测、大数据与网络安全、安全可视化与可视分析。

袁晓如（1975-），男，江苏吴江人，博士，北京大学研究员，主要研究方向为科学可视化、信息可视化、可视分析、计算机图形学和人机交互。