CCF大专委2017年大数据发展趋势预测

Developing tendency prediction of big data in 2017 from CCF TFBD

周涛(1979-),男,博士,启明星辰教授级高级工程师、大数据实验室副主任,主要研究方向为大数据安全分析、事件关联分析、入侵检测等。

潘柱廷(1969-),男,启明星辰教授级高级工程师、首席战略官,长期从事信息安全技术和战略研究工作。中国计算机学会(CCF)常务理事,CCF大数据专家委员会委员兼副秘书长,CCF计算机安全专家委员会常务委员,中国互联网协会常务理事,云安全联盟CSA中国区理事。

杨婧(1983-),女,博士,中国科学院计算技术研究所工程师,中国计算机学会大数据专家委员会秘书处工作人员。主要研究方向为数据科学、最优化查询处理、数据挖掘等。在SIGMOD、SSTD、DASFAA、APWEB、WWW Journal等一流国际期刊和国际会议上发表论文13篇。

程学旗(1972-),男,中国科学院计算技术研究所研究员、博士生导师、副所长,中国科学院网络数据科学与技术重点实验室主任。中国计算机学会理事、大数据专家委员会秘书长,中国中文信息学会常务理事、信息检索与内容安全专委会副主任。主要研究方向为数据科学、大数据引擎系统、Web搜索与挖掘、大数据安全。

中图分类号:TP399 文献标识码:A doi: 10.11959/j.issn.2096-0271.2017012

1 引言

2016年中国大数据技术大会（BDTC）于2016年12月8日在北京召开，在此次大会上发布了CCF大数据专家委员会（以下简称大专委）关于未来一年的大数据发展趋势预测，已成为每年大会的保留节目，也成为大专委的一项品牌活动。2017年的趋势报告一经发布，就引发了业界的广泛关注和持续传播。

本次大数据发展趋势预测经历了候选项征集和正式投票两个环节，最终形成的预测选项包括67项发展趋势选项和9项专项调研选项，投票范围面向大专委的正式委员和通讯委员。本次活动共收到有效投票82份，通过对这些投票的汇总、整理和解读，形成了本次对外发布的年度预测，这是大专委群体智慧的结晶。

任何形式的解读都难免掺杂解读者主观因素的影响。下面根据2016年的投票结果，对比往年的趋势预测情况，尽量从客观的角度对本次活动形成的趋势预测选项进行解读，以便读者对各个趋势选项进行理解。

2 2017年大数据发展十大趋势

2.1 趋势一：机器学习继续成为大数据智能分析的核心技术

在2016年的调查问卷中，将往年“深度学习继续成为大数据智能分析的核心技术”候选项的主题词，由“深度学习”调整为“机器学习”。做出这一调整的理由是深度学习只是机器学习领域的一个分支，虽然近年来在应用领域发展迅速，但其在算法层面上相对于传统神经网络算法并无本质上的创新；用机器学习取代深度学习，更能体现大数据智能分析的实质。

这一调整的效果立竿见影，该项得到了超过半数的投票专家的认可，使其成为得票最高的趋势选项。大数据的核心价值在于蕴含在大数据之中的、深层次的信息和知识，而如果没有机器学习技术分析大数据，对大数据核心价值的利用将无从谈起。随着机器学习与数据科学家们的关系越来越紧密，掌握机器学习的基础技能对数据科学领域的职业发展而言将成为一种必需技能。在大数据时代，依靠大数据管理和高性能计算的支持，机器学习将成为大数据智能分析的核心技术。

2.2 趋势二：人工智能和脑科学相结合，成为大数据分析领域的热点

近年来，随着深度学习技术在图像处理、自然语言处理等领域的成功应用，人工智能又迎来了新一轮的发展浪潮。2016年3月，Google公司的围棋AI软件“AlphaGo”战胜了顶尖职业棋手李世石九段，人类失守被视为“人类最后的智力骄傲”的围棋领域，引发了全社会的高度关注。此外，2016年恰逢“人工智能60周年”，在计算机领域的各项活动中人工智能都受到了广泛宣传。在此背景下，该候选项成为得票的榜眼就不难理解了。

脑科学是以人类大脑为研究对象的学科，在人工智能的研究中有一个流派，倾向于认为对大脑的恰当模拟会是制造出人工智能的关键。但事实上，直到今天，人们对人脑工作原理的了解仍然处于初级阶段，更谈不上对它的精确模拟了。目前，“类脑计算”最成功的例子就是深度学习，如果说基于对脑科学如此初级的认识，就产生了这样积极的效果，那么在对脑科学持续

研究的基础上进一步优化人工智能理论及应用，其效果就更值得期待。

2.3 趋势三：大数据安全与隐私令人忧虑

这是十大趋势中唯一影响大数据发展的制约因素。回顾历年的趋势预测可以发现，安全和隐私几乎每年都会排在第3、第4的位置。对安全和隐私的担忧使得数据开放步伐放慢，使得大家顾虑大数据的深度应用可能会带来负面价值，使得其成为大数据发展的重要威胁，这是需要解决的一个问题。

事实上，在信息系统建设中安全滞后于业务的发展，已经成为了一种惯例。系统的建设者首先要解决的是业务问题，在安全问题威胁到业务系统的正常运行时，才会考虑解决与之相关的安全问题。回顾大数据生态系统的发展历程，发现这种规律仍然存在。以Hadoop为例，早期的Hadoop版本中没有考虑对计算节点的认证、对数据的访问控制，而是仅仅考虑了分布式计算架构的核心问题。但随着Hadoop生态系统的逐渐完善，目前已经可以通过Kerberos对节点和用户进行身份认证；通过Apache Ranger和Apache Sentry对数据进行细粒度的访问控制；采用Apache Knox作为集群的统一安全访问入口；通过Apache Eagle进行异常行为检测和性能监控等。此外，对大数据的脱敏和匿名化技术也逐渐发展成熟。相信随着大数据应用越来越广泛、整个生态系统发展会越来越完善，对大数据安全和隐私的担忧可以通过技术手段逐渐化解。

2.4 趋势四：多学科融合与数据科学兴起

该选项同样是历年趋势预测的常客。大数据技术是多学科多技术领域的融合，大数据应用更是与多领域交叉融合，这种交叉融合催生了数据科学的产生和兴起。可以看到很多数据相关的专门实验室、专项研究院所相继出现，许多高校开展了以大数据技术及应用为特色的学位教育，数据科学作为一门新兴学科得到了持续发展。

数据科学的发展，反过来又促进了多学科的融合。许多学科研究的方向表面上看大不相同，背后却有相同的数据科学和计算科学的基础。例如医学和语言学是两个完全不同的学科，但如果在大数据的基础上借助人工智能实现智能诊疗和机器翻译，所采用的底层技术很大程度上是相通的。预期未来许多前沿学科的发展，都要依赖于本学科领域知识、数据科学与计算科学之间的融合。

2.5 趋势五：大数据处理多样化模式并存融合，流计算成主流模式之一

目前，大数据技术的应用已经从互联网行业逐渐扩展到各个传统行业，早已不再是什么“新”技术。作为一个可供佐证的论据，美国咨询公司Gartner在2015年底发布的新兴技术成熟度曲线中，首次缺失了大数据的身影。Gartner对此的解释是大数据的应用已经渗透到各行各业，应用模式也已成熟，不再是新兴技术。

由于大数据应用场景复杂多样，在不同的环境中，对数据产生速度、存储容量、一致性、实时性、容错性等方面的要求千差万别，很难有一个数据处理平台能满足所有场景的需求，这就导致了大数据处理模式多样化的现状。仅以NoSQL数据库为例，截至目前开源NoSQL数据库项目已超过225个，每个项目都有不同的数据处理模式。即使在同一个大数据业务系统建设中，为了适应不同的业务需求而综合采用

SQL、NoSQL和NewSQL等“混搭”的处理架构也已成为常态。

在众多处理模式中，流计算的发展格外引人注目。随着大数据应用场景对实时性的要求越来越高，流计算有可能超越传统Hadoop平台的批处理模式，成为主流的大数据计算模式。从开源生态圈来看，可供选择的流计算工具越来越丰富，Spark Streaming、Storm、Flink、Apex等工具快速迭代，Heron等新型工具不断涌现，Kafka也推出了流计算模块Kafka Stream；从架构来看，Kappa架构逐渐被接受，批处理成为了流计算的特例。

2.6 趋势六：数据的语义化和知识化是数据价值的基础问题

数据语义化和知识化是首次进入十大发展趋势预测的选项。数据语义化是指用特有的属性格式化文档内容，使得机器可以理解其内容；数据知识化是指挖掘和展示数据中的信息和知识。从Linked of Data的发展（目前为百亿三元组量级），到Google知识图谱和多种自然语言问答应用的出现，可以推断广大网络/移动用户在大数据时代获取信息时，越来越需要数据和信息的知识化组织和语义关联，也说明普通的个人用户从大数据中获得知识和价值是一个基本需求。

2.7 趋势七：开源成为大数据技术生态主流

大数据技术生态是伴随着Hadoop的开源起步的，预测开源会继续成为技术生态的主流形式。开源技术的蓬勃发展，大大降低了大数据的应用门槛，有力推动了基于大数据的业务模式在各行各业落地，也给传统数据管理厂商带来了严峻的挑战。

目前，大数据生态圈的发展势头迅猛，每当现有的技术不能满足新的应用模式时，总会产生多个与之相关的开源项目，从而带动新一轮的技术升级。在参与者方面，专业大数据企业、互联网企业、高等院校、科研机构，乃至某些政府机构和部门，都成了开源软件的贡献者。另外也看到，来自中国的开源软件及贡献者越来越多地进入了全球大数据生态圈，也促进了大数据技术在国内的发展。

2.8 趋势八：政府大数据发展迅速

政府大数据同样是首次进入十大发展趋势预测的选项，反映了大数据专家委员会的专家们对政府引导大数据应用的期待。政府的各级职能部门在日常管理中积累了大量的数据，特别是一些管理机构独有的、与人民生活密切相关的数据。笔者预测，一些拥有大数据的政府部门会利用积累的数据，采用大数据技术进行分析，提升社会治理水平。

另一方面，政府的数据是利用全社会的公共资源采集、存储和管理的，也是一种公共资源，也应当在经过必要的技术处理后为全社会所共享。目前，部分省市已经开放了多种类型的城市数据，中国政府已经有了数据整合的计划，相信这种趋势会越来越明显。

2.9 趋势九：推动数据立法，重视个人数据隐私

数据立法从另外一个角度上体现了专家们对大数据带来的隐私问题的担忧。人们在享受互联网上各项服务时，越追求服务的便捷化、个性化，就需要提供越多的个人数据，个人面临的数据隐私风险就越

严重。可以说，个人数据隐私与个性化服务之间本身就是一对矛盾。

单凭技术手段解决个人数据隐私问题是远远不够的，必须采用技术和管理并重的方式，这就需要推动相关的数据立法。要从数据的全生命周期进行综合考虑，从法律层面对数据的采集、传输、流转、交易、使用和销毁等环节做出明确约束，使得个人数据隐私保护有法可依。

2.10 趋势十：可视化技术和工具提升大数据分析工具的易用性

可视化也连续多年成为十大发展趋势预测的选项，2016年还占据了榜首的位置。2017年的投票关注度虽有所下降，但还是幸运占据了十大趋势的最后一席。

可视化是通过把复杂的数据转化为可以交互的图形，帮助用户更好地理解分析数据对象，发现、洞察其内在规律。它使得大数据能够为更多人理解、使用，使得大数据的使用者从少数专家扩展到更广泛的民众。笔者认为可视化技术，尤其是交互式可视化分析技术，是将人的直观分析能力与机器的强大计算能力相结合的有效方式，是提升大数据分析工具易用性的重要手段。

3 大数据发展专项调研分析

3.1 最令人瞩目的应用领域

大数据的发展最直接的推动力来自于应用，大专委5年来都做了“最令人瞩目的应用领域”的专项调研，结果见**表1**。这5年以来，互联网和金融一直是排在前三的领域。尤其近4年的调研，互联网、金融和健康医疗一直排在前三甲。2017年排在其后的城镇化和制造业等领域，得票数与前三相去甚远。而健康医疗超越金融排在第2位，其原因可谓多样。这也许和互联网金融及其所面临的风险有某种关联度，反映出专家们对应用领域的关注度有了一些微调。

3.2 在数据资源流转上会有什么举措

表2是从2015年开始做的关于“数据资源流转”的专项调研。做研究和应用都有一个切实难题就是“没有数据”“数据从哪里来”。可以看到，2017年的结果有一个有

表1 大数据应用最令人瞩目领域调查结果对比

序号	2013年	2014年	2015年	2016年	2017年
1	医疗	互联网、电子商务	互联网、电子商务	互联网、电子商务	互联网、电子商务
2	金融	金融	金融	金融	健康医疗
3	电子商务	健康医疗	健康医疗	健康医疗	金融
4	城市管理	舆情分析、情报分析	城镇化、智慧城市	城镇化、智慧城市	城镇化、智慧城市
5			社会安全、犯罪调查	舆情分析、情报分析	制造业、工业大数据

表2 数据资源流转举措调查结果对比

序号	2016年	2017年
1	会自己收集大量数据	会自己收集大量数据
2	会利用数据提供服务	会利用数据提供服务
3	会买数据集	会只下载和获得免费数据集
4	会只下载和获得免费数据集	会买数据集
5	没有数据流转考虑；会免费提供数据集；会提供数据流转和交易的中介服务；会卖数据；会进行国际数据交换	会免费提供数据集

意思的小变化，“会买数据集”退到第4位，而且票数和前面3项的差距也较大。这解释为当前大家对数据流转和数据交易并不抱太大的期望值，数据的流转并没有真正形成一个良好的环境和市场。因此，大家转而倾向于自己搜集数据、下载并获得的免费数据，而不寄希望于通过交易和价值交换带来有用的数据集，尤其对量较大的数据集的期望值在降低。当然，“会免费提供数据集”能够排到第5位，也让大家嗅到了一点共享的气息。

这条预测反映的问题其实还是比较严重的。数据只有流转起来才能产生巨大价值，促进并解决这一窘境，对整个大数据领域和数据领域来说极为重要。

表 3 与大数据最匹配的概念调查结果对比

序号	2016年	2017年
1	互联网+	智能计算或认知计算
2	云计算	云计算
3	智慧城市	机器人和人工智能
4	物联网、移动互联网	物联网、智慧城市
5	大众创业万众创新、工业互联网（工业4.0）、智能生活设备、一带一路	

表 4 我国大数据发展的最主要推动者调查结果对比

序号	2015年	2016年	2017年
1	大型互联网公司	大型互联网公司	大型互联网公司
2	政府机构	政府机构	政府机构
3	国内大学和科研院所	创业企业	
4	公共服务机构		
5	创业企业		

表 5 大数据发展阶段判断调查结果对比

发展阶段	2015年	2016年	2017年
极为初级	17%	33%	16%
即将快速增长	31%	40%	24%
爆发增长中	10%	9%	23%
达到一个顶峰，上升乏力	18%	4%	7%
达到一个顶峰，将下降和幻灭	5%	0%	4%
稳步成长中	20%	14%	26%

3.3 与大数据最匹配的概念

本项也是从2016年开始做的专项调研，见**表3**。从调研结果来看，与大数据最匹配的概念受2017年度热门话题影响较大。2016年专家们认为最匹配的概念是“互联网+”，2017年排名前三的选项中，有两项都与智能计算相关。很明显，大家对2017年的预测延续了2016年的热门话题“人工智能”。

3.4 我国大数据发展的最主要推动者

本项关注到底是什么样的力量在推动大数据的技术、产业、应用的发展，见**表4**。2015年调研结果是大型互联网公司、政府机构和大学科研院所等。在2016年做的第二次调研中，科研院所从名单中消失了，大家认为大学和科研院所并没有直接带来资本力量和各种资源。或许受到国家提出的“双创”概念影响，2016年创业公司成为一个大家寄希望的推动者，但2017年创业公司从这个名单中消失了。这个推动者名单中，只剩下大型互联网公司和政府机构，其他的票数都很少。这个专项调研结果与大数据发展趋势十大预测中第8条“政府大数据发展迅速”相符。互联网应用对大数据的推动力量是毋庸置疑的，真正新的推动力量只有政府机构这个选项。大家期望政府在治理理念上的变化、治理方法上的更新，带来对大数据应用的真实需求。

3.5 大数据发展阶段判断

本项借用Gartner技术成熟度曲线中对技术发展阶段的划分，评估大家对大数据的当前发展阶段的看法，见**表5**。从这6个阶段的投票分布来看，整体上大专

委的专家们对大数据的发展前景还是持乐观态度，也就是第2、第3和第6阶段占到主流，超过了70%的比例。而且有相当多的专家（26%）认为大数据已经过了幻灭期而进入稳定增长时期，也就是低风险而高价值创造的时期。虽然大数据的泡沫破裂期是不是真的度过，在大专委的专家中还有一定分歧，但总体来说对大数据前景的看好还是毋庸置疑的。

4 结束语

纵观本次2017年度大数据发展趋势预测，可以看到如下特点。

● 与纯粹技术相关的预测条目比例高。往年预测中有很多关于产业、资本、政策、生态、人才教育、职业等方面的条目，本次预测的十大趋势中，有7项是关于具体技术的，包括机器学习、人工智能、学科融合、处理多样性、流计算、语义化、开源、可视化等，这预示着专家们更寄希望于从技术本身寻求突破。

● 政策法规的动力和规范作用受到关注。从第8和第9条预测可以看出，大专委的专家们除了关注技术突破，还期望在政策和法规上能给予大数据发展有效的帮助。换句话说，大家在技术之外的发展诉求中，在市场力量和政府力量的天平中，更偏重于谋求政府的力量给予支持。当然，这也得益于最近一年政府方面在大数据方向上的明确态度，也得益于全国人民代表大会常务委员会对于《中华人民共和国网络安全法》的通过和颁布。

● 对安全和隐私问题的关注成为常态。可以看到，大数据安全和隐私问题条目每年都在预测中出现，在大数据实际应用中这也成为了严重的制约因素。对这一问题的解决，需要采用技术和管理相结合的手段。为此，一方面大数据生态圈的安全工具越来越丰富，另一方面对数据立法的期待也出现在了本次预测选项中。 □