第6篇

戴维·德维特访谈录

Interview with David DeWitt

戴维·德维特简介

戴维·德维特（David DeWitt），1970 年在科尔盖特大学获得学士学位。1976 年在密西根大学获得博士学位。同年 9 月份进入威斯康星大学麦迪逊分校计算机科学系任教并创立了威斯康星数据库研究小组，至今已经三十多年。1999 年 7 月到 2004 年 7 月期间，戴维·德维特任威斯康星大学麦迪逊分校计算机科学系主任。在并行数据库、测试标准、面向对象数据库和 XML 数据库技术上的创造性思维和开拓性工作，使他成为该技术领域公认的权威。戴维·德维特于 1995 年成为美国计算机学会（ACM）的会士，于 1998 年成为美国国家工程院的院士，于 2007 年成为美国艺术和科学院的院士。戴维·德维特因为对数据库领域的贡献，1995 年获得了 ACM SIGMOD 创新奖，2009 年获得了 IEEE Emanuel R. Piore 奖。戴维·德维特已经发表了 120 多篇论文。他曾任职于很多程序委员会和美国国家自然科学基金会审核小组。在 2000 ~ 2003 年之间，他任职于美国国家自然科学基金会（NSF）的计算机和信息科学与工程（CISE）顾问委员会。在 2005 ~ 2007 年之间，他任职于美国科学研究委员会的计算机科学与通信部（CSTB）。他还曾任职于几个核管理委员会和国防高级研究计划局的研究小组。2008 年从威斯康星大学退休之后，戴维·德维特成为微软公司的数据和存储平台部门的战略技术专家。德维特在威斯康星州的麦迪逊，与威斯康星大学麦迪逊分校的计算机科学系联合，为微软创建和领导微软吉姆·格雷（Jim Gray）实验室，这个实验室是微软的新的高级开发中心。

本专访主要介绍了戴维·德维特关于计算机学科课程的反思，以及什么是数据库研究团队应该引以为豪的，为什么查询优化不起作用，超级计算基金有时被拙劣地花掉，以及他为什么是一个不够好的编码人员，和不擅长去做数据库理论研究等。

玛丽安·温丝特

这期我们采访的是戴维·德维特（David DeWitt），时间是 2001 年的 11 月，在伊利诺伊大学厄巴纳 – 尚佩恩分校的罗伊·坎贝尔高清晰电视直播间。再一次感谢那些帮助我们为本次专访以及其他专访提供问题的人。

玛丽安·温丝特：今天，我们的客人是戴维·德维特，他是威斯康星大学麦迪逊分校计算机科学系 John P. Morgridge 教授，还是计算机科学系主任。他是美国国家工程院院士；还是 ACM 会士；[他是今年的 SIGMOD/PODS 会议的组织委员会主席]，并且他以性能评测和并行数据库方面的工作而著称。所以，欢迎戴维·德维特。

我喜欢从学术生涯开始问一些问题。您已经和工业界走的很近，在一个几乎没数据库企业的地方。你觉得数据库研究人员是应该去产业发达的地方工作，还是没有关系？

戴维·德维特：我不认为他们有太多的关系。我认为最重要的事情是挑选一个能够支撑起建立一个研究组的系。我认为在这样的学校可以更好地建立一个强大的研究组。我们 [在威斯康星] 已经展示了在美国中部可以很有成

效地做数据库研究。我认为在美东西海岸之外，还有其它地方有非常好的数据库组。

玛丽安·温丝特：讲到建立一个数据库组：大多数计算机科学系都有一到两个从事数据库研究的人，但是在威斯康星，许多年里，已经有 5 个或者更多做同类研究的人。随着人数的增加，是否会有一个质的改变，还是会和以前一样？

戴维·德维特：我认为一个组有 4 个或者 5 个成员会有一定的优势。如果你看威斯康星的系的设置，我们已经试着组织 4 ~ 5 个人一组。我认为两个人一起会出现一个周期性问题：时而他们不想一起工作，时而他们想一起工作。但是当这 5 个人是 Mike Carey、Jeff Naughton、我、Raghu Ramakrishnan 和 Yannis Ionnides，如果再加上 Miron Livny，会有很多种排列组合方式进行工作。所以，5 是一个有趣的数字，并且比 2 要好得多。总之，我认为除了量的不同之外还有质的不同。

玛丽安·温丝特：您已经领导并完成了学术界的很多较大的软件项目，这不是经常能做到的：这需要很多钱，很多人。您怎么样把精力合理分配到写论文和开发软件两类工作中？

戴维·德维特：我认为写论文和开发软件本来就是一个整体事件。我们先看一下我最近的两个项目，Paradise 项目和 Niagara 项目。Paradise 项目——如果以单个美元产生的论文数来衡量，这是一个很糟糕的项目。但我们开发了一个伟大的软件，也很高兴做这件事情；最多的时候有 25 人同时做这个项目，其中包括学生，也包括全职工作人员。对于学术界来说，这个项目太大了。发表的论文很少，单篇论文的花销确实很高。在 Niagara 项目中，没能开发一个可靠的软件，但是我们发表了很多论文。我认为你不能计划它；你只能顺其自然，有时你会创造出好的软件，有时你会发表很多论文，甚至有可能两个同时做到，就像在 Gamma 项目发生的一样。但是 Gamma 项目是

个例外。

玛丽安·温丝特：在那种项目中，论文和软件，哪一个更有影响力？

戴维·德维特：主要是依赖于着眼于什么。

玛丽安·温丝特：在这三个项目中，您着眼于什么？

戴维·德维特：在 Gamma 项目中，我们试着证明所提出的概念。其实 Naughton 和我成为了开发这个软件产品的最终的编程者。Jeff 设计了实验，而我编写了代码。我不是一个很好的编码者，而且直到我们写完代码，软件产品还不能用。

玛丽安·温丝特：那么它证明概念了吗？

戴维·德维特：它证明了概念。论文和软件哪一个更有影响力，主要依赖参与该项目的学生以及这些学生的能力。有时，你会有一些好的想法转化为好的软件，但是有的时候这些想法也可能不够好。我认为应该充分发挥学生的特长——他们软件开发的能力和他们做研究的能力。

玛丽安·温丝特：转换一下话题，对于计算机科学主修专业，什么内容应该加入到引导性课程中？

戴维·德维特：这是一个很好的问题，很重要的问题。实际上我们正在关注这个问题，因为如果你看到在计算机科引导学课程上的女生数量（至少在威斯康星）很少，你也会思考这个问题。问题是，为什么女生数如此少呢？

玛丽安·温丝特：它有多少？

戴维·德维特：大概在引导性课程上有 22% 的女生，但是主修的却下降到只有 10% ~ 15%。

玛丽安·温丝特：我所在的系比这还少。

戴维·德维特：为什么会发生这种情况？我不知道。我有两个女儿，一个主修化学，另一个主修数学，没有一个上计算机科学课程，虽然数学专业将来会用到计算机科学课程的内容。所以一定有某些事情使得高校女生不去上计算机科学课程，尽管这会使他们具有做计算机科学的完美能力。我不知道为什么会这样；我不知道是不是因为计算机科学被认为是男人主导的，或者被认为是枯燥的。问题的部分原因是我们先教编程。并且我认为对于大多数人来说，编程是枯燥的，编程不能完全体现计算机科学的发展。如果你思考化学，化学先从无机化学开始，定量分析只是引导课程中很小的一部分；引导课程的大部分还是无机化学。我们的引导课程可以先教他们一些体系结构、一些理论和一些数据库系统。但是不用一步到位直接到数据结构和编程。

玛丽安·温丝特：那么它们是否是实用课程？

戴维·德维特：或许是，或许不是。我不再认为编程是计算机科学的一个巨大部分。一定存在一些你可以做的计算机科学相关的事情，并且不需要你有太多的编程技能。我们应该从不同角度去尝试，看看它们是否会影响这个领域的女性数量。当然，这样做可能没有效果。

玛丽安·温丝特：那么您在威斯康星已经开始尝试这种新的引导性课程方法了吗？

戴维·德维特：没有，但是我们有一个课程委员会，主要由年青教员组成，他们刚刚获得博士学位，在最近的 25 年内我们没有改变我们的课程。引导顺序看起来还是和以前一样，什么课教什么内容。所以我们已经试着让这些低年级教员思考一些不同的方法，并且尝试这些新方法。我们将会在引导课程中尝试这些不同的方法，使其带有少量的编程，而把编程编入正式课程中。

玛丽安·温丝特：有趣。我期待听到好消息。

戴维·德维特：实际上，我希望其他人也思考并尝试一下，如果可以的话，我们会复制他们的课程。

玛丽安·温丝特：您可以作为我们其他人的先驱者。

很多年以前，您是其中一个最先流行的数据库测试基准的作者之一，测试标准是 Wisconsin 测试基准。关于这个测试基准，您是否有故事要告诉我们呢？

戴维·德维特：实际上，那是一个很有趣的经历。很多人都关注它。同时它也使得很多人（包括一些朋友）都很生气。我记得 Mike Stonebraker 就因为我很抓狂，因为我们测出了 Ingres 不能很好地处理一些特殊的查询。我认为很多人热衷于追求性能结果，而不是拿到结果做进一步分析，看系统哪里可行，哪里不可行。

这正是引起 Larry Ellison（注：Oracle 创始人）脑火的事情——我猜这是最好的故事——试图让我被解雇。他不明白任期的概念，也不明白系主任不会开除我是因为我不会说 Oracle 好话。其实，测试基准在圈内已经发挥了很好的作用。我认为这可以帮助开发者聚焦在他们的关注点上。总之，我认为整个测试基准的工作对于圈内的发展起到了积极的作用。

玛丽安·温丝特：您在暗示教授不应该去做测试基准，除非他们已拿到终身教职，是吗？

戴维·德维特：（笑）是的，我肯定建议这样！悲哀的事情是每个数据库产品 [我相信除了 DB2 以外] 都有一个条款声明可以说除了运营商之外其它人不可能发布数据，这都是 Wisconsin 测试标准惹的祸。我认为这很糟糕。这不是工业界应该有的态度。如果你卖一个产品，人们应该能够评估这个产品。数据库厂商似乎对人们评测其产品有恐惧感。

玛丽安·温丝特：但是厂商发布的测试基准结果通常也是经过独立审计的。

戴维·德维特：不是的，他们从来不审计。厂商在报告他们的测试基准数据的时候必须遵守许多规则，但是我认为大家普遍同意客户与厂商对评测的做法不同。

玛丽安·温丝特：这就是说：厂商发布的测试结果一定是不可超越的。

戴维·德维特：那是当然，那一定是上限。

我认为这个条款使得厂商只侧重某一数据，如 TPC-A 或者是 -B 或者是 -C 或者是 -D 或者是 -H，总的来说这伤害了本领域或用户的利益，因为用户不可以自行评测并发布其结果。

玛丽安·温丝特：可以发布，只要你把数据库系统称为 A，B，……

戴维·德维特：……C 或者 D。是的，这是标准托词，但是仍然奏效。

玛丽安·温丝特：对于学术界，创业热（Startup Fever）是一件好事还是一件坏事？

戴维·德维特：坏的一面是一些好的学生不再愿意读 PhD，好的一面是学术研究可以有更好的经费支持。我认为总体上来说它是中性的。我不得不说它已经伤害了博士的质量。

玛丽安·温丝特：如果创业热降温后会怎样？

戴维·德维特：我认为很好。人们会更愿意坚持下去读一个博士学位、学生们可能会不再读完硕士后就奔向工业界，可能会更多考虑留在学术界发展。短期内这对学术界有好处。

玛丽安·温丝特：在美国，最近的经济出现了衰退，您觉得这对学术界会有什么样的影响？

戴维·德维特：学术界也在遇到同样的事情。研究生院会收到更多的申请，因此未来能招到更好的学生。这就意味着我们将培养出更多高质量的博士生，并且有望使得更多的学生对学术感兴趣并且继续从事学术研究。

玛丽安·温丝特：可是学术研究经费能否负责起这些学生？

戴维·德维特：我认为现实问题是政府经过 9·11 事件之后是否还有能力资助那些需要资助的项目？并且是否会对基础的研究产生不利的影响？我认为如果你是做安全研究的，那是正当时，做数据库系统研究也不过时，因为政府需要管理的信息在增加。随着政府收集信息日益增加，数据库系统和信息管理会变得日益重要。同时也出现了隐私问题，这是我们不得不担心的。我认为这对于数据库研究团队是一件好事情。

玛丽安·温丝特：继续前面的数据库经费问题的讨论。我知道您是美国国家科学基金会（NSF）的计算机和信息科学与工程（CISE）顾问委员会的成员，并且 CISE 是数据库研究方面的 NSF 资助的主要来源，就像其他研究领域一样。您认为 NSF 应不应该资助人，或者资助一些特殊的研发项目？

戴维·德维特：我认为他们应该资助尽可能多的项目。有时资助一些人也是很好的。有时项目建议书内容比较窄。但是我认为需要能够资助新的教员，所以有时你需要资助一些项目申请。但是资助一些人也是非常必要的的。

我不认为 CISE 顾问委员会对 CISE 所做的决定会有什么的影响，所以人们不要指望我能为获得资助说上话。

玛丽安·温丝特：那么您给了他们什么建议？

戴维·德维特：不管我们对他们说了什么，他们从来不听取我们的建议，所以没关系。我不确定为什么 CISE 有一个顾问委员会，因为我认为我们的建议被一次又一次地忽略。

玛丽安·温丝特：您说您认为 NSF CISE 应该资助更多的项目，但是您

也说您认为一些项目申请书的内容面太窄。

戴维·德维特：如果你说你想做X方面的工作，并且X实际上很宽泛，我认为很难得到项目的资金支持。一个典型的策略是做深入的研究，然后写一个申请书——我认为这是不合理的。我们应该允许人们在一个比较宽泛的题目下做研究，这正是我支持资助人的考虑。

我认为整个资助情况，即使是NSF的ITR，都很令人失望。在过去的一段时间里，有一个项目叫做协同实验研究（CER）；政府从20世纪70年代后期开始，每年花掉100万美元左右，并且你能用一部分钱去做一些重要软件开发。现在，你可以得到的最大的ITR资助是每年100万美元——并且20年已经过去了！现在获得的每年100万的投入产出要比过去的还要少。我认为这真的很不幸。从我个人角度来看，我认为CISE把太多的资金投入到了超级计算机、兆级和网格计算差不多都给了UIUC。

玛丽安·温丝特：确实我们是此类资助和研究的试验品。

戴维·德维特：我认为那种资助没有资助计算机科学；他资助的是物理学家，并不是资助计算机科学家。

玛丽安·温丝特：我获得了资助。我的安全方面的工作，你刚才提到过……

戴维·德维特：好好。我认为此类资助付出了太多的钱。我认为构建2000个结点的集群并且声称它是计算机科学，这些都是废话。

玛丽安·温丝特：您不想它们能够模拟原子裂变吗？

戴维·德维特：我认为那是在资助物理学家，而不是资助计算机科学研究。

玛丽安·温丝特：哦。好，我认为为了能够模拟原子裂变，他们需要很

多帮助，因为写这种模拟器很困难。

戴维·德维特：我认为很多花在超级计算机上的钱被浪费掉了。我认为PACI项目就是一个很好的没有充分利用资金的一个例子。

玛丽安·温丝特：您是被采访者，您问这个问题，不考虑我的感受。

戴维·德维特：不，那很好。你不觉得PACI有些事与愿违，或许其整个方向就是……

我认为资助大型设备（比如说超级计算机）很好，因为你需要具有国家级的计算中心，比如伊利诺伊州[NCSA]，以及匹兹堡[超级计算机中心]和圣地亚哥[超级计算机中心]；我认为你需要有超级计算机中心，在那里人们可以脱离政府实验室来完成他们的计算。但是我不认为应该绑定设备资助和研究资助。那是我对PACI的一点自己的看法。我认为它就是在试图绑定设备资助和研究资助及应用资助，但我认为他们应该是三个独立的部分。实际上我比较喜欢匹兹堡的模式，把设备资助和研究资助独立分开，而不是作为一整块进行总体资助——因为我认为这样更合理。

玛丽安·温丝特：我本打算问您为什么，可是您已经告诉我并且更好地说明了。

戴维·德维特：资助机构应该承担更多的责任。

玛丽安·温丝特：所以您的意思是说，举个例子，如果他们在构建大设备时成功了，他们可能称整个项目取得了一个成功，即使……

戴维·德维特：我不认为在构建大型设备的时候有什么研究工作。你可能只需要买一大堆机器，堆到计算机房，然后把它们聚在一起，把他们连接成网格。我只认为购买硬件属于资金使用范畴。很明显，如果你买硬件，那么你就应该支持维护硬件；但是你应该不必需要获得授权的人来决定哪个研究项目应该被资助。我只是不喜欢那种模式，这就是为什么我退出PACI项目

的原因。

玛丽安·温丝特：我知道，很有趣。

戴维·德维特：我不认为 SIGMOD 同仁们会对这个感兴趣。

玛丽安·温丝特：好吧！只是我对它比较感兴趣。那是我生活中的全部。[对于 SIGMOD 同行，它似乎显得很无聊，] 我们可以立即从访谈录的打印版里删除这部分。

数据库研究的传统核心领域不再像以前一样被资助。是否这就代表我们这个领域比较成熟了，或者说，是否我们已经错过了一些需要更多研究的核心领域？

戴维·德维特：我认为我们已经错过了一些需要更多研究的核心领域。

首先，我认为这个领域已经很成熟了。我们现在已经有很能干的系统，并且这个领域也应为取得这样的成就而自豪。我认为学术研究者和工业界人士都做出了突出贡献。这些系统即可靠又可扩展，同时提供很高的性能。我认为我们就这个领域已经做了一个相对较完美的工作，并且每个人都应该因为这个而自豪。

但是，我认为有很多核心领域需要更多的关注。查询优化就是一个很大的漏洞；同时我认为 I/O 也是一个大漏洞。我认为很多人已经进入这些热门领域。一会儿是递归查询处理，一会儿又是面向对象数据库，一会儿是数据立方体；因为吉姆·格雷写了一篇很好的关于数据立方体的文章。那之后，我们发现有 300 多人写了关于数据立方体的文章。现在我们还有数据挖掘，KDD 会议就有 700 多人参加。我认为人们着迷于那些热门领域——这很好，因为我认为只有一小部分人对核心问题感兴趣。

对于核心数据库研究，只有很少的资金资助。美国国防部先进研究项目局（DARPA）已经对这个不感兴趣很多年了；DARPA 现在也对数据库没有投入，虽然这可能改变，并且 NSF 也不感兴趣，所以几乎不大可能获得资助做

这些核心研究。

玛丽安·温丝特：您说查询优化需要很多研究，那么查询优化的哪方面需要更多的工作？

戴维·德维特：整个查询优化！查询优化已经有 22 年的历史了。每个人都在做同样的事情，所有的工作都是基于 Pat Selinger 和 System R 团队所做的工作，但效果并不好。数据库系统已经变得很强大。现在，我们的数据库系统用户可能要做 10 路连接，我们可能在可扩展机器上的大数据集之上运行 TPC-H 查询（具有难以置信的复杂度的查询）。如果没有手动调优，想为这些查询产生可靠的较好的计划，查询优化器实现起来就会表现得很糟糕。我认为我们需要重新考虑怎么进行查询优化，因为数据库其它技术已经得到很大提高，而查询优化却没有得到提高。

玛丽安·温丝特：对于我们应该怎么样做查询优化，您是否有什么特殊的建议？

戴维·德维特：我的想法是我们应该采用 Ingres 的查询处理机制，它基本上采用在优化和执行阶段进行迭代进行的方式。现在则是先优化数据库操作，然后执行。我们完全是基于数据统计的荒谬假设来优化九路和十路连接查询计划。现实是经过数个查询之后，我们就没办法预测有多少个元组会被查出来。你不知道连接列的属性值是否有关系；你不知道你的直方图是否还准确——或许你根本就没有直方图。所以，查询优化器在处理查询树中有 5、6 层的连接时会做一个理想化的假设。

我个人观点是我们需要重新审视一下我们该怎么做优化和执行。现在，我们先优化然后执行。取而代之，我认为我们需要观察一些事情，比如说，优化一点，执行一点，优化再多一点，执行再多一点。我们应该从不同角度去尝试，因为这是一个没有技术提高的领域。

但是，这不代表说 Pat Selinger 在她的工作里没有做出巨大贡献。当你写

一篇可能结束这个领域的论文时，明显地，它就会是一篇超级论文，Pat 就是一个超级巨星！但是现在我们能在执行方面做点什么，我们需要回去重做查询优化。而只添加更好的直方图不能够解决问题。我不知道怎么做，但是这是一个我认为很重要的方向。

玛丽安·温丝特：当您指出查询优化和 I/O 时，您的意思是通过 I/O 解决吗？

戴维·德维特：我的意思是磁盘变得越来越慢。如果你实际地看过传输率，磁盘是变快了；但是如果你让容量除以传输率，你会发现磁盘实际上是变慢了。

一些人建议应该把 SQL 处理器放在磁盘控制器中，创建一个智能磁盘。我认为那不会帮助我们解决问题；我认为智能磁盘看起来只是像一个旧的数据库机——处理器和磁盘绑在一起。

在威斯康星，我们正在试图寻找一种解决问题的方法：我们试图看一下是否能够做一个虚拟分片工作。这是一个很旧的想法；MCC 的 Bubba 项目做过这个想法，并且把它叫做分解存储模型。这个想法是，如果你只需要一个表的一个或者两个或者三个或者多个列，为什么你要读整个表？垂直分片能够使得硬件缓存得到充分利用；它使得压缩更容易实现；它可能极大地增加你所用的 I/O 设备的效率。

很明显，作为数据库人，我们不能去改变磁盘的制造过程。我们不得不与商业磁盘共存。并且最近几年他们已经达到了半 T，两年之后会出现 1T；到 2010 年会出现几个 T 的磁盘。数据库不会像磁盘一样增长那么快，除非你处理图像或者视频。

总之，我只认为 I/O 是一个大问题，并且现在厂商只生产不管问题，因为磁盘变得越来越便宜。或许有一些 I/O 方面的问题会让我们感兴趣去做一做。

玛丽安·温丝特：还有其他什么领域，您想提醒大家需要额外关注的吗？

戴维·德维特：肯定有其他领域，但是这两个是我现在正在考虑的。

玛丽安·温丝特：您有没有喜欢的热门领域？您是否愿意看到人们追逐热门？

戴维·德维特：很明显，XML 是一个热门领域。我认为对 XML 感兴趣的原因是数据库研究者在研究分布式关系数据库时遇到了失败，并且我认为 XML 是很灵巧的，因为如果确实可行，并且人们使用 XML，其网站运行 XQuery，那么你就可以考虑构建一个巨大的分布式系统。我认为那是令人兴奋的研究领域，数据库界正在研究它。我认为在大规模上解决分布式数据库系统问题，对我们来说，在接下来的几年间会成为一个有趣的挑战。但是已经有很多人在研究 XML 和 XML 数据库。

XML 数据库不是一个核心领域，但是我猜测它是一个热门领域。并且会引出新问题：我们能否考虑语义，和人工智能同行一起处理这些内容？单有 XML 是什么也做不了的，为了能够智能地处理大量数据，所以需要集成其他技术。

玛丽安·温丝特：在数据库界，很多人都有一种感受，学生发表了比以前多的改进性论文，因为如果它是一个改进性论文就很容易发表到顶级会议上去，因为很容易处理在改进性论文中审稿人提出的所有漏洞，并且学生不得不比过去发表更多的文章以找到一份较好的工作。真的是这样吗？改进性论文真的越来越多吗？如果是这样，是否这也意味会有问题？如果有问题，我们该怎么解决？

戴维·德维特：我不确定有越来越多的改进性论文。

我认为有一个基本问题，就是 SIGMOD 和 VLDB 论文评阅方式。我最近写了一篇论文，被 SIGMOD 拒绝但是被 VLDB 接收并被评为最佳论文。论

文在两次投稿过程中基本上未做修改。现在，就有了这种错误，如果一篇论文被一个会议拒绝，但是同一篇论文被另一个会议认为是很好的工作。

我不知道要怎么调解处理。我认为论文接收或者不接收变成了一个随机事件。我认为我们需要引进一个反馈调解处理的循环，你首先应该投稿，编委会会审阅你的稿件并且给你一些反馈意见，然后给你一次机会去反驳它，直到编委会满意；或者我们需要多轮的处理。

我认为现在让论文被会议接受的过程完全是一个掷骰子的过程。我认为这对年青教员来说很难。作为一个资深教员，当我的论文被拒收我也会感到慌张，这是真的，尽管论文是否被接收和我的工作前景没有关系！特别是由于我是我们系的系主任，所以院长设定我的薪水并且我不由我的同事来审阅，并且院长也不会看我是否有两篇 VLDB 拒接的论文。但是对于一个不是终身教职的年轻人来说，自认为是不错的论文被拒，理由又不太清楚，就是很不爽的事情。

玛丽安·温丝特：您提到的这个方式听起来很像期刊的审阅方式。您是说让 SIGMOD 转化成 TODS 一样的？

戴维·德维特：当然不是，因为这些年 TODS 除了理论论文没有别的。

玛丽安·温丝特：我希望不是这样。

戴维·德维特：期刊处理是开放的，但是会议的程序委员会不是开放的。现在这个时间线完全是荒谬的。我们第一次提交论文在 11 月，在隔年六月发表。粗略计算一下之间有 8 个月时间。我们都知道论文已经录入计算机。从快照准备（复制）到生产的整个处理过程不是一个事件。有一个很长的窗口期，从 11 月 1 号到 3、4 月份，在这段时间我们可以执行调解处理。它不像一个期刊，因为它只有一轮审阅和讨论。你提交你的论文；你从审阅者那里获得意见；你有一个机会写反驳意见给评阅人；并且你不需要修改你的论文。然后由程序委员会处理。

我建议这种方式是因为我认为，程序委员会成员有时因为不能很好地了解这个领域而评阅这篇论文，或者他们会误解作者意图。我认为我们应该试图去尝试改变，因为在处理论文是否应该被接收的过程中存在着太多的不确定性因素。

我还认为我们应该录用更多的论文，可以超出论文报告数。一些人作了很好报告；一些人作了不怎么好的报告。举例来说，SIGMOD 有 250 篇论文投稿，可能收录 75 到 100 篇成为一个论文集，然后只挑出来 25 ~ 30 篇论文做会议报告。我认为没有必要每一篇被录取的论文都报告。一些论文会更合适作报告交流。

玛丽安·温丝特：当您挑出 25 ~ 30 篇论文，您怎么知道您挑出来的都有最好的报告？

戴维·德维特：我还真不知道。我刚才在想，让我们做一些改动！就像引导性计算机科学课程那样：我们已经做了很久类似的事情，自 1979 年我一直参加 SIGMOD，到现在它都没有变化——让我们做点什么让它改变。

玛丽安·温丝特：如果 SIGMOD 有现在的两倍大，即接收的论文是现在两倍，这会有帮助吗？这是否会让录取过程变得随机性越来越小？

戴维·德维特：如果 SIGMOD 一年增长两倍，我认为会有帮助，或 VLDB 明年去比香港更合适的地方，实在太远了。

玛丽安·温丝特：好吧！但举办地对生活在香港的人们来说是很合适的。

戴维·德维特：是的，对于生活在香港的人来说是公道合理的，但是对于生活在美洲和欧洲的人来说是不合适的。基于现在的系统来说，一年组织两次 SIGMOD 会议是很困难的，因为当前我们安排每一届 SIGMOD 会议在不同地方。大公司有行业展示，他们雇用人帮助他们运作行业展示。做

SIGMOD 和 VLDB 的程序委员会委员不会那么困难。困难的是如何处理所有当地的组织安排。我认为我们有足够专家资源使得我们可以在美国每年召开一次额外的会议。

玛丽安·温丝特：听起来很有趣。

对那些初出茅庐或处于职业生涯中期的数据库研究者和从业者，您有什么建议吗？

戴维·德维特：我认为我的建议与我给那些年青教师的建议没什么不同。（作为系主任，我不得不操心这些事情。）我认为重要的是挑选一两个方向，并且做出很好的工作。我认为一个年青教师可能做得最坏的事情就是泛而不精。如果你想做数据挖掘，好，那就努力成为一个数据挖掘方面最厉害的人之一。不要试图做数据挖掘、数据立方体、XML 和内存数据库。挑选一两个方向，把所有注意力都集中在挑选的方向上。

我的其他建议是不要太早地带太多的学生。我认为一个年青教师的学生数量最多的时候不应该超过 3 ~ 4 个，学生是很好的资源，但是如果你有太多的学生，你就不能以很简单有效的方式和他们进行工作。

玛丽安·温丝特：您通常有多少学生？

戴维·德维特：很多了！我现在有 7 ~ 8 个，并且我在试图回到 3 ~ 4 个的水平。

玛丽安·温丝特：7 ~ 8 个博士研究生吗？

戴维·德维特：大部分是博士研究生和少数几个大学生。我开始越来越多地指导大学生。

玛丽安·温丝特：他们有时可能有用。

戴维·德维特：他们很有用。

玛丽安·温丝特：如果您在工作中可以做一件应该做而现在还没有做的事情，那会是什么事情？

戴维·德维特：我没办法回答你这个问题……去游泳池好好地游泳？

玛丽安·温丝特：作为一个计算机科学研究者，如果您能够改变关于自己的一件事情，那会是什么？

戴维·德维特：我希望我有强有力的数学背景知识。我认为有很多东西我不明白，但是又希望弄明白。我本科学的是化学，所以没有上那么多的数学课程。我认为这使我不能参与很多的研究工作。我猜这是我希望可以改变的一件事情。

玛丽安·温丝特：如果您有这个背景知识，您是否会做更多数据库理论方面的工作？

戴维·德维特：可能吧！我不可能做这种工作。我不够聪明去做数据库理论工作。我有一个 PODS 论文，有时人们会拿那个取笑我。但是那个是我学生的论文，不是我的。

玛丽安·温丝特：非常感谢您参加我们的访谈！

戴维·德维特：谢谢你们邀请我！

（范玉雷 译，孟小峰 审校）