

格哈德·葛惟昆访谈录

为什么我们应该迎接挑战？为什么说SQL语言功能太强大了？精确度只是一个传说？如何在德国组织大型研究团队以及更多内容。

玛丽安·温斯特

欢迎来到ACM SIGMOD record数据库杰出人物系列访谈，我是玛丽安·温斯特，现在我在VLDB2007的主会场维也纳和格哈德·葛惟昆在一起，格哈德是位于德国赛布魯班的马克思·普朗克信息研究中心的研究主任，同时，他是ACM院士，曾获得VLDB总统特别奖、SIGMOD最佳论文奖、VLDB的10年优秀论文奖以及CIDR永恒构思奖，他在达姆施塔特大学获得博士学位，欢迎格哈德。

玛丽安：我们刊物的大多数读者和观众都不怎么认识德国教授，您是在马克思·普朗克研究中心的唯一的数据库教授。从美国人的观点来看，计算机系的任何一个领域都不是以人多制胜的。你是如何在这种体系下组织了一个如此优秀的研究团队？

葛惟昆：最需要注意的一点是，你周围必须有优秀的人。优秀的学生是一笔宝贵的财富。如果你的团队由优秀的学生组成，较少的人就可以做出很好的工作。

我来稍微纠正一下你对德国体制的观点。在德国，只有到了相对成熟的阶段才能成为一名教授，然后转为终身教授。我们正在逐步的采用美国体制，比如，先成为一名终生助理教授。当然，研究者不能在研究生毕业之后消失十年，然后再做一名教授。在成为教授之前的中间阶段，研究者可以有博士后、研究助理等头衔。我也有几名这样的研究人员。

我现在在马克思·普朗克研究中心，同时我也为赛不鲁班大学工作，这两个机构在同一个校园。所以，我在大学里面也有同事，以前的同事是克里斯·托夫科赫，从2008年秋天开始，我又同傑生·迪特里希一起工作，他是从苏黎世联邦理工学院过来的，和我走的同样的路，我是15年前来的。傑生是一个非常优秀的同事，我们已经开始合作了。

玛丽安：你现在涉足不同的研究领域，起初，您只是研究事务处理，是什么让您涉及到这么多的研究领域呢？

葛惟昆：是好奇心驱使的。我想学习新东西、理解新东西，最好的途径就是在新领域内做研究。它会促使你去理解某些问题的研究现状、研读文献。我一般从教这个科目开始研究新领域。比如，当我7年前转向信息检索领域时，是由于我要开始教信息检索的课程。通过教课可以得到研究思路。

当我回首过去的时，我的研究生涯可以大致分为三个部分。第一部分是事务处理。（顺便说一下，那时候我们并不认为研究事务处理是狭窄的）然后，在90年代初期，我开始研究自动调谐，持续了大概10年的时间。在2000年左右，我转向数据库以及信息检索集成领域。

玛丽安：您不认为这些年来您的研究领域变得宽泛了吗？

葛惟昆：确实变得宽泛了。因为这三部分在时间上有重叠。每当开始一个新研究时，我并没有完全停止前面的研究工作。但是有些重叠并只做原创性的研究。比如，在自动调谐课题上，最近我开始和Surajit Chaudhuri做一个tutorial。这并不是原创性研究。Surajit仍然在做原创性的研究，我在教育方面也加入到研究中去。

我现在的学生比原来多了，由于在马克思 普朗克研究中心，必须建立一个大的研究团队才能以多制胜。在加入马克思 普朗克研究中心之前，我在大学里和很少几个学生一起工作，我认为这样更加专注。现在，我们一个组就有20个研究人员，其中10-15个人是学生，其余的是博士后或者在美国被称为年轻的助理教授一类的人员。我倾向于给有天赋的学生一些灵活性。如果他们自己有好的想法，或者他们能在某个问题上说服我，证明他们有足够的能够解决这个问题，我就会让他们自己去。我们的一些研究工作（比如，机器学习）就是由一个学生的构思促进的，然后我就以指导者的身份加入了这个研究。

玛丽安：如何防止自己被分的太细了呢？

葛惟昆：这是一个很好的问题。你必须一直监督自己，看看自己是不是足够专注。我尝试着把一两个课题作为主要的课题，围绕着这些课题一些搭建系统。我们并不是为了单纯的搭建系统而去做系统，系统的搭建是我们研究工作聚焦的点，也是所有子课题集成的点。如何让学生彼此充分交流是最大的问题，因为他们从不同的角度研究同一个系统，他们要知道自己做的部分和其它部分如何交互。我在大学里和5-10个人一起工作时，我们通常有一个面向系统的工作再加上一些半成熟的工作，这样做是冒险的，但是对未来的研究有很大的帮助。现在，在马克思 普朗克研究中心，由于团队规模大了，我们有两个以系统为中心的工作。我要同时扮演两个角色：所有项目的研究负责人以及其中一个项目的研究人员。

玛丽安：你从需要精确研究的领域转向了一个事事模糊的研究领域，是什么让你做出这种转变的？

葛惟昆：我认为精确的概念是一个传说。当年，在对薪资数据做工作时，确实是很精确的，我的薪水，当然希望它是精确的。但是，当谈论到科学数据，比如测量数据时，实验给你一个值，通常是有误差的，在某种程度上是不确定的。我们只是忽略了这种不确定性。我们假定温度是精确的35.5789度，但事实上并不是这样。在文本世界中，或者是在文本和数据结合的领域中，将文本以及用户所需要的信息结合起来本身就含有不确定性。对于研究信息检索的人来说，一个查询的结果是对用户需求的无限趋近，并不是一个精确值。

玛丽安：听起来您已经做好充分的准备去面对现实了。

葛惟昆：确切的说，一个人在充满文档的世界中不可能做到完全精确，就像从文本中抽取实体和关系

一样。最重要的是做这件事有多大的置信度。就像现在转向了数据聚集、数据组织，并且在这些数据上进行查询，最终很可能偏离你、原来的起点，从而导致对查询结果的误解。所以，最重要的是获取置信度或者非置信度，在每一步都需要研究置信度如何传播到下一步。我对概率数据深信不疑。

玛丽安：在不同领域的工作经历能让您告诉我们数据库领域的走向是什么？未来的研究点在哪里？

葛惟昆：显然，这是非常主观的，我的答案会和我的研究课题结合起来。我看到处处都有信息爆炸的情况。在我看来，文本是人类产生效率最高的一种信息，因为产生精确的数据，也就是有模式的结构化数据需要太多时间了。文档的核心的主要内容不过是数据库中的一个元组罢了。但是大多数人写一篇评论比在数据库中建立一个描述这个评论主题的元组容易。所以，我期待看到文本数据量的持续增大，比如，含有标签、书签的Web2.0，博客，公告牌等等这类数据。

人们的交谈比写文本更加高效，语音识别是更好的方向。目前研究潮流是关于谈话处理的，从谈话中抽取有趣的元素，比如，谈话者的激动情绪，沮丧情绪或者言语带讽刺性等等内容。

玛丽安：是的，但你不能告诉我语音识别是数据库领域的研究。

葛惟昆：我不管你把我放在哪里，我是一个计算机科学家。我认为我们应该看看其他领域在做什么。我不是在说语音识别时数据库领域的研究内容，但是它有一部分是关于数据管理的，还有一些其它部分。

玛丽安：有些人认为，您从精确问题转向了统计问题是“面向黑暗的诱惑”，并结束了人工智能的研究，你认为这是这样的吗？

葛惟昆：确实是这样的，我不会反对这种说法。当我年轻的时候，我也想过这样一个谚语“被黑暗诱惑”Mike Stonebraker曾经说过：“这个问题是人工智能完全问题”，也就是说，这个问题是没有希望解决的，是科幻小说中的，是和“把我传送上飞船”类似的。当我年龄稍大一些时，我认为这种观点是错误的。

我们应该主动寻找挑战。物理学家经常自己主动寻找挑战性的问题，就像受控核聚变一样，他们在很多领域有大型研究项目，如等离子物理学。他们并不信守承诺，他们说在三十年内有重大突破，但是并没有达成。这就告诉我们，他们的问题是很难的。在生命科学中，科学家不仅仅满足于研究染色体的构成，也想研究染色体的语义、功能等，这是世纪级别的科研项目。所以，我们计算机科学家应该成长起来，应该更加自信的寻找挑战。我们可以把挑战分成一个个步骤，显然，如果30到50年内我们看不到任何进展也是不合理的。如果我们给自己定一个目标，百年之后有人认为研究是正确的或者是歪曲的，我们是活不到那个时候去验证对错了。因此，我们应该把大挑战分成不同的里程碑，分成一些小的、短期的目标。大的挑战还放在那里，等待我们继续完成。

玛丽安：告诉我们你在2002年获得VLDB 10年优秀论文奖的一些事情吧。

葛惟昆：那是我们在1992年发表在VLDB上的一篇文章，那年的VLDB在温哥华召开。我们的论文是关于锁管理中的负载均衡的问题。在高负载情况下，可以快速进入锁竞争环节，比如，当有混合负载时，负载急剧攀升的情况。我们的工作是关于准入控制和并发控制的，所以，它是非常狭窄却非常有技术含量的。我不认为是论文本身帮我们拿到这个10年优秀论文奖的，关键是它背后蕴含的一个主题：自动调谐。

1989年，我结束博士后工作后来到了奥斯汀的MCC，在苏黎世联邦理工学院做助理教授。在MCC，我们在攻克一个较大的难题——大型并行数据库机器，在当时这是很大的问题。我受到这个项目的影 响，想做一些非常有挑战性的、长期的研究工作，那时，很少有人研究比较偏僻的问题。我是年轻的助理教授，我也没有太多的可用资源。

我开始研究通过向系统中加入更多智能元素进行自动协调的问题。我们继续采用有适应力的方法，我们采用了随机模型以及反馈控制，也就是这篇VLDB1992论文的重要基础。我认为当委员们把我们这篇论文选为2002年的10年优秀论文奖时，他们承认了我们所做工作的重要性：自动调谐、自我管理 系统或者是现在被称为自动计算的技术。

玛丽安：说说你在2005年CIDR会议上获得的永恒构思奖的事情吧。

葛惟昆：这是一个趣味奖项，和我获得的其它奖项有很大不同。在我的网页上，我把它和许多正式奖项列在一起，因为我认为人们不能过高的看待正式奖项。在获得这些奖项时，有时候是靠一点运气，有很多优秀的人该获得这些奖项可是现在还没得到。

CIDR的这个奖项是在深夜里颁奖的，人们把它叫做“Gong Show”。候选者有20到25个人，每个人5分钟的时间，随便讲点什么内容：挑战、未解决的问题、他们喜欢的问题、甚至笑话，什么都行。人们喝了点酒，在晚上10点时很放松，我的报告是半正式化的。我的主题是在我的研究领域中的实验方法。显然，这是一个正式的话题，但是报告本身是非常有趣的。我给大家讲了一些如何把你的实验结果展示的更好的技巧。我的实际要点是我们应该认真的把实验原则反应出来。和其它领域相比，我们确实只是小孩子。我的报告的幻灯片可以从网站下载，顺便说一下，最后的一两页幻灯片上有一些值得认真思考的问题。这是很及时的报告，它结合了最近SIGMOD对已录用的论文的实验可重复性的要求。

玛丽安：你怎么看待SQL？

葛惟昆：作为一种语言？我从来没有研究过如何设计一种语言或者其它关于语言的工作。但我有大量的教学经验。当我教SQL的时候，我想给出一些正式的语义，比如，转换到关系代数或关系演算。如果只对SQL的子集这样做事很简单的，比如选择，投影，连接。但是如果对于复杂的SQL语句来说，比如有5级嵌套子循环，几个左外连接，有关系的变量、聚集或者分组操作，几乎没有学生能把这个查询写对。我们用实例来教这样的问题：老师用新的例子展示如何使用这些新特性，希望学生能知道如何

正确使用新特性，但是学生们做的并不好。

我认为这种教法会让我们的学生今后写出非常糟糕的SQL语言，有很多错误。但是对SQL语言进行调试并不是简单的工作，因为它是公开的。一种语言越公开就越难调试。调试SQL语言要求我们把复杂的表述分成简单的SQL块。但是，人们为什么需要功能如此强大的语言呢？

所以，我认为用一种功能更少的语言好。我的榜样是Pascal语言的发明者Niklaus Wirth，他的语言中有“少就是多”的哲学。这种语言架构简单，属性较少，易于使用。语言的特性太丰富是祸根。我知道这是有争议的，很多人都不同意，但这是我的观点，没有别的意思。

玛丽安：语言越高级就越难调试，那是不是说汇编语言是最容易调试的？

葛惟昆：这是一个很好的意见。就状态驱动的调试而言，当你看到程序的状态是真的，或者接近真的。但是易于调试并不是语言的唯一重要的方面。设计、编码、注释所需要的时间也很重要。

我不是说高级语言是错误的。我是在说如果想变得高级，语义必须非常精确，必须教给大家这些语义。但愿学生们能从几个简单例子中理解这些超级难的语义。

玛丽安：TODS上有一篇论文给SQL语义做了形式化的阐述。

葛惟昆：用那篇论文来教学太难了，顺便说下，这不单单是SQL的问题，是所有高级语言的问题。我认为把这些语言特性定义的太丰富了是有问题的。

玛丽安：你现在是马克思 普朗克研究中心的主任，还是VLDB基金的主席，你还组织其它机构吗？

葛惟昆：噢！我不想。服务性的工作总要有有人来做，于是我就做了，也算为计算机领域做点事。我年轻的时候，参加会议也享受会议，但总有人要组织会议，举办会议的机构也需要人来组织。我只是尽了自己的责任。

玛丽安：你在管理上、会议组织上都做得很好。现在人们的要求越来越高了，您如何平衡做服务的时间，做管理者的时间以及做研究的时间呢？

葛惟昆：我觉得我在这些事情上做的并不比别人好。每当我接受一个工作时，我都尽力把它做好。这是我的态度，也许在我的基因中。但我并不是一个天生的管理者或者主席。你说我在这些事情上做得好，其实我只是花费时间尽力做好这些事而已。

我正在努力控制参加服务性工作的时间。我现在还欠数据库领域一些债，也许到某一个天，我认为我的债已经还完了，我就会拒绝类似的工作。到那个时候，也许我会再写一本书什么的。

玛丽安：谁终结了并行和分布式信息系统会议（PDIS）？

葛惟昆：是Jeff Naughton。Jeff和我是1994年迈阿密召开的最后一届PDIS会议的程序委员会主席。事实上，我是在开玩笑，结果我和Jeff就终结了PDIS会议。实际上，PDIS是自我终结，这没什么不对的。

现在也有像PDIS一样的专题会议，如，演绎及对象数据库系统（DOODS），也有一些workshop

在逐年召开。这些专门会议和workshop的主题当年都是未被主流会议如VLDB, SIGMOD覆盖的。从提交的论文还是参加者方面看, 当他们走向主流后, 主流会议就吸收了这些方向, 这时, 结束这些专题会议也是合乎潮流的, PDIS正是如此。

玛丽安: 我知道你的意思, 虽然主流会议上有很多投稿, 看起来我们仍然有必要开些好的专题会议。

葛惟昆: 现在情况不同了。比如, 网络空间正在爆炸, 那么就应该有一个关于web 2.0的会议。当然, 整个领域也有了新的会议。与典型的数据库人不同, 我也在关注我们周围的领域正在做什么, 信息检索和web上的研究。比如, 有个叫WSDM的新会议, 是关于web上的搜索, 挖掘等研方向的, 也包括Web2.0等内容。

玛丽安: 听说你曾经一个人去沙漠度假, 你认为这个假期怎么样?

葛惟昆: 我喜欢沙漠。我在美国度过了一段非常有意义的时光, 我喜欢西南地区。我确实单独一个人背包旅行。我没有提前设计这次旅行, 但有时候你在一个空间停止之后不能再回溯了。我在西南的南犹他州呆了一阵。有时候你会遇到一些障碍, 你可以继续前进但不能后退, 因为你已经习惯于遇到困难了, 就像是水滴石穿一样。

玛丽安: 如果这次旅行很危险, 你认为一个去是明智的吗?

葛惟昆: 没有比我一个人走更危险的事了。我的人生中曾受到几次伤害, 有一次我在人行道上等行人通过, 结果造成了踝关节韧带断裂, 所以说没有比我独自行走更加危险的事情了。

玛丽安: 确实是, 然而即使是踝关节韧带断裂, 周围仍有一些人可以帮助你。如果你在沙漠中人迹罕至的地方, 周围没有人, 手机也可能没有信号。

葛惟昆: 确实是这样的。实际上, 在这几次旅行中, 我没有带手机, 因为我很清楚那些地方没有信号。一般的情况是, 先在某些机构进行注册, 如美国土地管理局。当快到返回的日子时, 他们会开始搜寻。在这几次旅行中我比较谨慎, 不会去冒不必要的风险。

玛丽安: 你认为欧洲基金组织对大型项目的影响有哪些?

葛惟昆: “大型”意味着很多不同的意思; 你可能是指项目中参与者数目比较多。所谓的大型项目有很多种类型。一个项目有20个参与者或者更多, 就是网络型, 这个项目有可能被划分为若干个5个人一组的小组。在小组之间的联系可能比较松散, 也就是人与人之间的交流不多, 小组之间的联合开发目标也不够明确。这类的大项目还不错。我所参与的人数较多的大项目都是网络型的。

一些人数较多的项目在项目结束时需要硬性的、可以交付的系统。如果要非常严格的按照官方的目标来衡量这些项目, 很可能都是失败的。我并不相信20个参与者可以做联合开发, 特别是一半以上的参加者是学术合作者且预算有限。

也有一种项目我们称之为STEPs, 尤其是研究型项目。这种项目大概有5-10个参与者。其中有3-4

个技术人员及研究人员，其它参与者可以做技术转换或者需求分析。当有些公司承担这些项目时，他们可以扮演不同的角色。公司可以承担开发或技术类的主要任务，也可以提供一些用例、需求或者潜在的商业模型。后者的角色通常在项目的起始及终止阶段起作用。这样可以将项目的参与者减少到数量较小的一些核心研发人员，这种团队非常有效。

玛丽安：你对无经验的人或者实习的人有什么建议吗？

葛惟昆：在计算理论领域我认识的人不多，有很多年轻的刚起步的人，也有一些40岁左右的人。在欧洲，如果你是一个理论学家，并不出名，已经快40岁了，也没有终身教授的职位，那么，你就会比较麻烦。有很多25岁的新星。这些40岁的研究者也并不差。他们有一些熟练的技巧，也可能考虑转移到其他领域去施展自己的技能和方法。反过来说，我认为对研究者在实际应用方面的要求应该更高一些。然而在天赋和理论之间的比例有一些失衡，因为理论更加具有挑战性。有一些人并不擅长理论，但是转移到实际应用领域会做的很好。

玛丽安：在你之前发表的论文中，有没有你最喜欢的工作？

葛惟昆：有一些我非常喜欢并且引以为傲的工作。有一个不太知名的工作是在WebDB2000上的workshop上发表的。我们做了在XML上的排名检索工作，这篇论文只有6页，题目叫做“Adding Relevance to XML”。那时候，研究XML的浪潮已经过去了，都是研究模式的。我认为没有方法将整个世界都统一到一个模式中。针对异质的、多样的、模糊的数据，就需要排名。这篇论文也是我们在数据库和信息检索结合领域的第一个工作。

玛丽安：如果你现在有时间在工作中再做一件事，你想做什么？

葛惟昆：和Surajit Chaudhuri合写一本书。

玛丽安：关于自动调谐？

葛惟昆：你说对了。

玛丽安：作为一个计算机科学家，如果你可以在某一个方面改变自己，你想改变什么？

葛惟昆：我想在年轻的时候多学一些理论—数学和其它理论。这些东西非常有用，当你有足够多的基础知识时，做研究会非常有效。

玛丽安：非常感谢您今天接受我们的访谈。

葛惟昆：谢谢你，玛丽安。

(霍峥译，张金增校)