

克里斯托斯 访谈录

玛丽安·温斯特

本专访主要介绍了幂率法则、有关分形、数据挖掘的未来，以及以及休假等话题。

玛丽安：欢迎来到计算机协会数据管理专业组对著名数据库社区成员的系列采访，我是玛丽安·温斯特，今天卡内基梅隆大学的一位计算机科学教授克里斯托斯也在这。他于 1989 年曾获得美国国家科学基金会（NSF）授予青年科学家的最高奖“青年研究者总统奖”，1997 年以他的 R+树论文荣获 VLDB 十年论文奖，1994 年以有关在时间序列数据库中快速子序列匹配的论文获得 SIGMOD 最佳论文奖。克里斯托斯是 SIGKDD 国际会议执行委员会成员，他对数据挖掘，数据库性能，空间和多媒体数据库有着广泛的兴趣。他从加拿大多伦多大学获得博士学位。欢迎克里斯托斯！

克里斯托斯：非常感谢玛丽安邀请我来到这。

玛丽安：你被称为合作大师，你不仅在数据库方面帮助同事，而且在其他学科如科学，统计学对来工业来访者有帮助，对于培养这些合作你都做了什么呢？

克里斯托斯：谢谢你的恭维，我认为对合作的渴望是我的一种个性。有人更愿意在一个领域做深入研究，也有人愿意合作。我曾经非常有幸能有卓越的工业，统计，机器学习领域的同事，这些合作是由自身发展起来的，我并没有特别的来培养。

玛丽安：你是怎样学习这么多领域如此多技术的基础知识的，并把它们运用在解决数据库问题上？

克里斯托斯：我有个准则：如果我遇到一个被运用两三次的方法，或者一个被重新改造两三次的方法，那这个方法很可能运用在数据库方面。分形体就是一个例子，现在我们正努力用同样的方法研究单值分解以及独立组件分析，因为这些技术对很多领域都有潜在的深远影响。

玛丽安：听说你是“全世界最好的人”，“无私的人”，因此我可以肯定这些称号跟你成功的合作是有关系的。有人建议我询问你如果做到成功的微笑！

你两个同样是计算机科学界的兄弟，你跟他们一起写了被称为“Faloutsos cubed paper”的论文，那是一个有关幂法则对 1997-8 的网络拓扑非常有影响力的论文。这些幂法则是什

么，如今还有效吗，我们为什么在数据库社区应该考虑它们？

克里斯托斯：幂法则在当今仍然起作用。我有些曲线图显示幂法则在 1997-8 月后一直有效，而且不仅仅是在计算机网络方面！就像齐普夫定律：有些词经常出现，而大多数词很少出现甚至从没出现过。对于网络连接也是一样：有些结点很受欢迎。每个人都想链接到 AT&T,IBM 或 Sprint,而没人愿意连接到很小的 ISP。对于公司的规模也是如此：很多大公司拥有二十五万员工，但是绝大部分公司仅有一两个人。因此这些幂法则不仅在论文中提到的网络上起效，在其他环境依然有效，并且几世纪以来都在起作用不仅仅是最近。

玛丽安：我们应该在数据库领域的哪些地方运用幂法则呢？

克里斯托斯：我们可以在选择性估计中，柱状图中运用到幂法则。Yannis Ioannidis 在 2003 年就以他的关于柱状图的论文获得了最佳论文奖。就是因为齐普夫分布才使得柱状图非常的成功：如果你保存少量最重要属性的频率数，那么其它的属性就无关紧要了。

幂法则跟分形关系密切。幂法则，分形，自相似在很多机器装置中出现，我们可以用自相似来解决维数灾难问题。在数据库和数据挖掘中，如果存在很多属性，我们就称有了维数灾难问题。如果存在过多的属性，那么大多数数据挖掘算法的运行时间会以指数级增长，因此高维数就成为问题。但这并不是维数决定的，而是一种数据集的分形维数（内在维数）决定的。通常分形维数都很低，这是因为属性重要性的偏态分布。可能存在上百个属性，但最重要的仅有几个，因此问题就没有我们想象的难了。

玛丽安：这样说的话，分形才是关键，那么新的最重要的问题又是什么呢？

克里斯托斯：分形对于很多问题的解决都很有用，比如图与社会网络的分析。目前已经有几乎所有种类的图的幂法则存在，自相似很可能也是。社会网（一个人认识另一个人），生物网，食物链网（一个吃另一个），这些都有自相似与分形。对于我们正在研究的传感器时间序列也有自相似。时间序列具有猝发性，自相似可以很好的描述这种猝发性。比如一个人会有一段沉默期，然后一段爆发期，更大的一段沉默期，更大的一段爆发期，与偶尔才有一次这样事件发生的标准泊松分布刚好相反。不，应该说这些事件非常的聚集和自相似。所以我认为分形可以解决很多问题而不止一个。

玛丽安：数据挖掘领域是一个很新的很有生机且发展很迅速的一个领域。你认为数据挖掘的未来的主要方向是什么，或者说这个领域将通向哪里？

克里斯托斯：这又是一个很好的问题。这绝对在很多方面都存在可能，比如网络，计算机网，

社会网，生物网，调控网络；这里强调了网络。对于时间序列分析也是如此，因为我们将面临很多来自传感器的测量，并且我们想从这些测量中得到模式。如果我们有了一个网络，我们就想找到其中的干扰，因此我们就会测量我们没个时间单元会得到多少数据包或检测。生物信息学也应该成为一个热点领域，实际上它已经是个热点领域了。

玛丽安：你是说数据挖掘这个领域将暂时致力于应用领域？

克里斯托斯：对，但这只是我个人片面的看法，因为我更多的面向实用领域。我有很多数据挖掘方面有理论思想和面向统计的同事，我确定他们有不同的观点。他们将很期待数据挖掘在数学问题上的深入研究。所以说我的看法只是从应用领域来说。

玛丽安：对于社会网，我们将挖掘哪种模式，我们在寻找什么？

克里斯托斯：我们想得到一个一般模式，像韩家炜关于艾滋病毒分子所做的那样，他是想弄清楚哪些子分子活跃。而我们是想知道在一个公司中哪些人活跃，或者弄清楚一些群体对一个部门是具有破坏性还是建设性，我们也想知道哪些是离群的边缘。比如说，我们有一群通常相互不说话的 researcher，如果我们在他们之间存在边界，那这些边界会是重要的界限。这些边界要么是令人怀疑的，因为这不应该发生，要么是很有价值，因为这些边界将成为使个部门协调工作的桥梁。对于数据挖掘的问题我们不是在寻找什么的特别的，而是寻找我们还不知道的一种能帮我们压缩这些数据集的模式。

玛丽安：数据挖掘领域包括所有统计学,人工智能,数据库背景的人。我听说来自一个领域的人不能够听懂其他领域的人的会议讨论。我也听说学统计学的人，在跟你一番谈话之后，他总是可以提出一些能打破你索引结构的疑问。于是我就想问拥有索引的关键是什么，由于存在这些相互不了解的分支学科的联盟，数据挖掘这个领域将发生什么呢？

克里斯托斯：我认为将要发生的事现在已经在发生着：很多会议把这些来自不同领域的人聚在一起，前几年会不太容易，之后大家相互了解彼此的心理，这正是我们现在做的。在数据库课堂上我们教卡方分布测试，因为它对于学统计学有帮助；我也肯定学统计学的人也在教二叉树索引。我不太记得你所提到的具体情况，但现在已经有很多学科交叉了。确实前几年可能会比较困难，但为了最终的目的这些困难还是值得的。

玛丽安：似乎很多希腊人都研究数据库，你对这有什么看法，这是一个机遇还是一个负担或者其他的什么？

克里斯托斯：我认为这是一个令人高兴的巧合。是二八定律和分形在起效，是一种聚集效应。在我读本科时一些数据库教授回到希腊，像 Dennis Tsichritzis，当然他们对数据库非常有激情。后来我们都去了美国或加拿大，然后同样的情况重复了好几次，一批一批的数据库研究人员出现，创造了指数性增长。现在我们有希腊学生和教授在研究数据库。我确定在其他国家也有同样的例子；像印度和以色列也有很多数据库教授。因此我说这是一个令人高兴的巧合。

玛丽安：你第一次来卡耐基梅隆大学是作为休假的游客，后来成为教授留在了那里。从马里兰大学（一个基于数据库的学校）转到耐基梅隆大学，在你到来的时候还没有一个数据库院系的成员，当时是什么情形？你当时又是怎么处理必须向身边的人证明你所学的整个学科？

克里斯托斯：实际上那是一个令人愉快的转变，因为当时卡耐基梅隆大学很积极创建了数据库组。的确我不得不做些证明，但那也主要是我所应该做的：教育。我必须告诉人们数据库不仅仅是信息的集合，它是一个有结构化查询语言的表的集合。在卡耐基梅隆的人都很好而且很愿意跨学科学习；当时通过招聘过程人工的挑选跨学科人员。所以这样使我的任务变的很容易。

玛丽安：你当时是怎样说服他们相信数据库很重要？你是否争论数据库具有经济上的重要性，或者只是从智力上感兴趣？

克里斯托斯：我一点都没必要去争论，因为他们已经很确信数据库研究非常重要，所以他们在邀请了我。

玛丽安：可是 Natassa Ailamaki 说她必须一遍一遍地证明她的学科。

克里斯托斯：我们当时是有很多要解释不是证明，因为他们有大量的数据而且想要处理这些数据。甚至当初作为游客的时候，他们就说：“你懂数据库呀，太好了！我有些问题，你能帮我吗？我有猴脑的时间序列，但怎样存储它们呢，又怎样寻找它们的相似处呢？”，他们想研究神经生物学，想弄清楚当人们用刺激物刺激猴时，猴脑是怎样运行的。所以说当时不是证明的问题，是一个速成课问题。

玛丽安：我听说你休了不少假期，并且你喜欢以特别的方式度假。那么对于考虑休假的人们你什么推荐吗？

克里斯托斯：我认为在一个工业实验室休假很有价值，因为假期可以帮助我们接触到实际问题实际客户，并且能接触更深的知识。所以我休了两个假期，一次是跟Rakesh Agrawal还有Bill Cody一起在IBM度过，另一次是跟Avi Silberschatz 和H. V. Jagadish在AT&T度过，当时他们俩都还在AT&T。因此我的建议是努力发现实际的客户需要什么。

玛丽安：要想接触实际的客户，你不需要去一个发展集团吗？

克里斯托斯：实际上不需要。因为我们的合作者与客户有直接或间接的联系。跟客户关系密切但又不是极其的密切，不是跟客户面对面的采访，但他们的抱怨需求会最终影响到研究实验室的。

玛丽安：你对没有经验的职业生涯中期的数据库研究者或从业者有什么建议吗？

克里斯托斯：我对他们的主要建议就是请享受他们正在做的事情。如果你发现一个话题很有趣，那么其他人也将会发现它有趣的。在得到终身教职前，当然遵守游戏规则挺重要的：如果学校需要杂志刊物，我们要确保有适量的。

得到终身教职以后，我认为人们就可以很自由的做自己最喜欢的事了，这是一个个人口味的问题。就我个人而言，我更喜欢研究有实际重要性并且能运用一些好的理论的问题。可能别的人更喜欢专注于实际问题；只要问题对公司或社会重要，那么他们就研究它。还有些极端的，有些人研究纯理论的问题不管这些问题是否有实际应用。我认为这三种都是有价值的，人们应该追求能让他们兴奋的东西上。

玛丽安：在你得到终身教职以后，你就可以自由的做你喜欢的事了，可是你还有那些学生呀。他们为了得到一个工作，很多人都不得不做很多事像教授助教写很多论文一样。你是怎样处理这些枯燥无味的事呢？

克里斯托斯：我觉得人根本不能摆脱枯燥无味的事！很不幸人们总认为得到终身职务后就可以放松了，其实不是，没人会放松的。可能只是思想上更自由更平静了，但仍然有很多工作要做。我想这个工作量读大学时和工作时都是一样的。这只是一个状态和心理的问题。

玛丽安：我们不应该告诉学生们这些，是吗？当他们读了这些不会沮丧吧？

克里斯托斯：我认为不会的。我们说的这些都是事实，他们都很聪明而且他们也看到了教授们（包括助教，同事，全职的，以及退休的）每天工作 10 到 12 个小时甚至更长时间。但这是个很有趣的工作，人们很享受去做，所以我认为这并没有不好。

玛丽安：这就很好的引出我下一个问题：假如你有足够多额外的工作时间去做其它的事，那你会做什么事呢？

克里斯托斯：没什么特别的。

跟你的工作差不多的吗？

对，差不多：画些草图，收集一些数据，找些模式，找找下一个能解决之前提到问题最好的工具。

玛丽安：在你的以前的工作研究中，你最喜欢的是什么？

克里斯托斯：应该是 94 年有关如何使用分形表征大量不均匀的点的论文，这样我们就能弄清楚 R 树的性能以及其它空间存取方法。

玛丽安：作为一个计算机科学研究人员，如果你能改变你自己，那你将改变什么？

克里斯托斯：可能是更加有条理，现在我的事情不那么有条理。

玛丽安：有时候教授让他们的秘书或博士后整理事物，这样就可以有条理，你以前试过吗？

克里斯托斯：还没有，这是个好主意，我应该试试。

玛丽安：我听说在希腊数据库社区你很爱讲笑话，你能讲个笑话来结束我们的采访吗？

克里斯托斯：当然了，我知道的最短的笑话是：我是个无神论者，谢谢上帝！

玛丽安：非常感谢。

克里斯托斯：谢谢你的访问。

（张啸剑译，马友忠校）