第16篇

克里斯托·法拉特访谈录

Interview with Christos Faloutsos

克里斯托·法拉特简介

克里斯托·法拉特（Christos Faloutsos），希腊人，1987年在加拿大多伦多大学获得博士学位，目前是卡内基·梅隆大学的计算机科学教授。1989年他曾获得美国国家科学基金会授予的青年科学家最高奖“青年研究者奖”，1997年他以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。他对数据挖掘、数据库性能、空间和多媒体数据库的研究做出卓越的贡献。

本专访主要介绍了幂率法则、分形方法、数据挖掘的未来，以及休假等话题。

玛丽安·温丝特

玛丽安·温丝特：欢迎来到本期 ACM SIGMOD Record 数据库领域杰出人物访谈。我是玛丽安·温丝特，今天卡内基·梅隆大学的一位计算机科学教授克里斯托·法拉特（Christos Faloutsos）接受我们的访谈。1989 年他曾获得美国国家科学基金会授予的青年科学家最高奖“青年研究者奖”，1997 年他以 R+ 树论文荣获 VLDB 十年论文奖，1994 年他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。克里斯托是 SIGKDD 国际会议执行委员会成员，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！

克里斯托·法拉特：非常感谢玛丽安·温丝特邀请我来到这儿。

玛丽安·温丝特：您被称为合作大师，既有数据库界同事，也有其它学科如科学界、统计学界以及工业界的合作，您是如何做到这样广泛的合作？

克里斯托·法拉特：谢谢恭维，对合作的渴望是我的个性。有人更愿意在一个领域做深入研究，也有人愿意合作。我非常有幸能与工业界、统计学界、机器学习领域的卓越伙伴进行合作，这些合作是由自身发展起来的，我并没有特别地有意为之。

玛丽安·温丝特：您是怎样学习这么多领域中的技术和基础知识的，又

是怎样把它们运用在解决数据库问题上？

克里斯托·法拉特：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，那这个方法很可能可以运用在数据库方面。分形方法就是一个例子，现在我们正努力用同样的方式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·温丝特：听说您是“全世界最好的人”，“无私的人”，因此我可以肯定得到这些称号跟您的成功合作是有关系的。有人建议要咨询您是如何一直保持微笑的！您有两个兄弟，同样在计算机科学界，您跟他们一起写了被称为“法拉特立方”的论文，该论文描述了有关幂率法则（Dower law）问题，并对 1997-8 版的互联网拓扑有着深远的影响。这些幂率法则是什么，如今还有效吗，如何在数据库领域应用？

克里斯托·法拉特：幂率法则在当今仍然起作用。有些数据曲线显示，幂率法则在 1997 年 8 月之后一直有效，而且不仅仅是在计算机网络方面。就像齐普夫定律（Zipf's Law）：一些词经常出现，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互联网服务提供商。对于公司的规模也是如此：很多大公司拥有 25 万员工，但是绝大部分公司仅有一两个人。因此这些幂率法则不仅在论文中提到的网络上起效，在其它环境依然有效，并且几世纪以来都在起作用而不仅仅是在最近几年。

玛丽安·温丝特：应该在数据库领域的哪些地方运用幂率法则呢？

克里斯托·法拉特：可以运用于选择性估计以及柱状图中。亚尼斯·伊奥尼迪斯（Yannis Ioannidis）在 2003 年就以柱状图为主题而获得了最佳论文奖。而这恰恰是因为这些齐普夫分布体现了柱状图的特点：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。

幂率法则跟分形关系密切。幂率法则、分形、自相似性在很多机器装置中出现，我们可以用自相似性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，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。但这并不是总体维数决定的，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，但是只有最开始的 3 ~ 5 个是最重要的，因此问题就没有我们想像得那么难了。

玛丽安·温丝特：这样的话，分形才是关键，那么新的问题又是什么呢？

克里斯托·法拉特：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，几乎所有种类图的幂率法则都存在，自相似很可能也是。社会网络、生物网、食物链网，这些都有自相似性与分形的特性。目前我们正在研究的传感器时间序列也具有自相似性。时间序列具有突发性，自相似性可以很好地描述这种特征。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，这与偶尔才有一次这样事件发生的标准泊松分布刚好相反，应该说这些事件非常聚集和自相似。所以，分形技术可以解决很多问题。

玛丽安·温丝特：数据挖掘领域是一个很新的、很有生机且发展很迅速的领域。您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·法拉特：这又是一个很好的问题。数据挖掘技术可以应用于很多方面，比如 Web、计算机网络、社会网络、生物网络以及调控网络等；而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，并且想从这些测量结果中得到模式。如果有了一个网络，我们就想找到其中的干扰，由此就会测量每个时间单元会得到多少数据包或脉冲信号（ping）。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·温丝特：您是说数据挖掘这个领域将暂时致力于应用领域？

克里斯托·法拉特：对，但这只是我个人片面的看法，因为我更多的研究是面向实用领域的。很多数据挖掘方面有理论思想和面向统计的同事，他们有着不同的观点。他们非常期待数据挖掘在数学问题上的深入研究。所以，我的看法只是从应用领域来说的。

玛丽安·温丝特：对于社会网络，将挖掘哪种模式，我们在寻找什么？

克里斯托·法拉特：我们只是想得到一般的模式，像韩家炜教授关于艾滋病毒分子所做的那样，他是想弄清楚哪些分子活跃。而我们是想知道在一个公司中哪些人是活跃的，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的，也想弄清离群之间的边界。比如说，有一群通常相互不说话的研究者，如果在他们之间存在边界，那这些边界会是重要的。这些边界要么是令人怀疑的，因为这不应该发生；要么是很有价值的，因为这些边界将成为部门协调工作的桥梁。对于数据挖掘的问题，不应该寻找什么是特别的，而是应寻找还不知道的一种能帮助我们压缩这个数据集的模式。

玛丽安·温丝特：数据挖掘领域包括了许多有统计学、人工智能、数据库背景的人。我听说某一个领域的人不能够理解其它领域的人的报告。我还听说一位统计学者对你说，你的每一个报告之后他都可以找到一个查询打破您提出的索引结构。那么索引的关键是什么？那些相互不了解的分支学科的介入，会使得数据挖掘领域发生什么？

克里斯托·法拉特：将要发生的事已经正在发生着：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。在数据库课堂上传授卡方分布知识有助于理解统计学；而统计学老师也会在课堂上教 B- 树索引，具体的细节我记不太清楚了，但现

在已经出现了许多交叉学科。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·温丝特：似乎很多希腊人都从事数据库研究，您对这有什么看法？这是机遇，还是一种责任，或者还有其它的综合原因？

克里斯托·法拉特：这确实是种令人高兴的巧合。也是二八定律和分形技术起的一种聚集效应。在我读本科时，一些热衷于数据库研究的专家（像 Dennis Tsichritzis）回到希腊。后来我们都去了美国或加拿大，类似的情况前后出现了几波，涌现出一批数据库研究人员，这种现象呈指数级增长。目前有很多希腊学生和教授在从事数据库研究。其他国家也同样，像印度和以色列也有很多数据库教授。因此，我说这是一种令人高兴的巧合。

玛丽安·温丝特：您第一次来卡内基·梅隆大学（CMU）是作学术休假，后来成为教授留在了那里。是什么促使您从马里兰大学转到卡内基·梅隆大学，要知道卡内基·梅隆大学在您到来之前还没有一个从事数据库研究的教员，当时是什？您又是怎么向身边的人证明您的学识？

克里斯托·法拉特：实际上，那是一个令人愉快的转变，因为当时卡内基·梅隆大学确实想创建数据库组。是的，我确实要对自己的学科做些证明，但是我主要从事的是数据库教学。我让大家知道数据库不仅仅是信息的集合，它也是一个有结构化查询语言的表的集合。卡内基·梅隆大学的人非常友善，有跨学科的氛围。其聘用流程就体现跨学科，所以让别人了解自己的学识很容易。

玛丽安·温丝特：当时您是怎样说服他们相信数据库很重要？

克里斯托·法拉特：没必要去说服，他们已经确信数据库研究非常重要，因此邀请了我。

玛丽安·温丝特：可是 Natassa Ailamaki 说她必须反复地证明她的学科。

克里斯托·法拉特：当时是有很多方面需要解释而不是去证明，因为他们拥有大量的而且想要处理的数据。甚至我当初作为访问学者的时候，他们就问："你懂数据库，太好了！我有些问题，你能帮我吗？我有关于猴脑的时间序列数据，但怎样存储呢，又怎样寻找它们的相似性呢？"，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，猴脑是怎样反应的。所以说当时数据库不需证明，而是要成为大家的速成课。

玛丽安·温丝特：听说您休了不少学术假期，并且喜欢以特别的方式度假。对于考虑休假的人，您有什么建议吗？

克里斯托·法拉特：我认为在一个工业界实验室做学术休假很有价值，因为这可以帮助我们接触到实际问题和实际客户，并且能接触更深的知识。所以我休了两次假期：一次是访问 IBM，与 Rakesh Agrawal 和 Bill Cody 一起度过的；另一次是访问 AT&T，与 Avi Silberschatz 和 H. V. Jagadish 一起度过的，当时他们俩都还在 AT&T。因此，我的建议是努力发现客户的实际需求是什么。

玛丽安·温丝特：要想接触实际的客户，您不需要去一个开发组吗？

克里斯托·法拉特：实际上不需要。因为我们的合作者与客户有着直接或间接的联系。与客户的关系是密切的，但又不是极其密切，不是与客户面对面的交流，但他们的意见需求会最终影响到实验室的研究。

玛丽安·温丝特：对没有经验的且处于职业生涯中期的数据库研究者或从业者，您有什么建议吗？

克里斯托·法拉特：对他们的主要建议是请享受自己正在做的事情。如果你发现一个话题很有趣，那么其他人也会发现的。在得到终身教席之前，遵守晋升的游戏规则是很重要的：如果学校需要期刊论文发表，要确保自己有适量的论文数目。

得到终身教职以后，你就可以很自由地做自己最喜欢做的事了，这与一个人的研究兴趣有关。就我个人而言，我更喜欢研究有实际价值、并且能运用的一些好的理论问题。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，他们就研究它。还有些极端的人，他们研究纯理论的问题不管这些问题是否有实际应用。这三种方式都是有价值的，研究者应该追求能让他们感兴趣的问题。

玛丽安·温丝特：您提到在得到终身教职以后，可以自由地做自己喜欢的事，可是您还有那些学生。他们为了得到一个工作，不得不做很多事（像当助理教授，写很多论文）。您是怎样处理这些枯燥无味的事呢？

克里斯托·法拉特：人无法摆脱那些枯燥无味的事！人们总认为得到终身教职后就可以轻松了，其实不然，没人会轻松。可能只是思想上更自由、更平静了，但仍然有很多工作要做。工作量与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·温丝特：不应该把这些告诉学生们，是吗？他们读了这些不会沮丧吧？

克里斯托·法拉特：不会的，这些都是事实，他们都很聪明，而且也看到了教授们（包括助教、同事、全职教员以及退休教员）每天工作 10 ~ 12 个小时甚至更长时间。这些都是很有趣的工作，我们很享受去做，并没有什么不好。

玛丽安·温丝特：这帮我引出了下一个问题：假如有足够多额外的时间去做其它的事，您会做什么事呢？

克里斯托·法拉特：没什么特别的。

玛丽安·温丝特：跟您的工作差不多的吗？

克里斯托·法拉特：对，差不多：画些草图，收集一些数据，找些模式，针对之前提到的所有问题，试着弄清楚什么是最适合使用的工具。

玛丽安·温丝特：在以前的工作研究中，您最喜欢的是什么？

克里斯托·法拉特：应该是 1994 年有关如何使用分形表征大量不均匀点的论文，那篇论文能帮我们弄清楚 R 树的性能以及其它空间存取的方法。

玛丽安·温丝特：作为一个计算机科学研究人员，如果您能改变有关自己的一件事情，您将改变什么？

克里斯托·法拉特：可能是做事情更加有条理性，目前我做事情不那么有条理。

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

克里斯托·法拉特：还没有，这是个好注意，我应该尝试一下。

玛丽安·温丝特：听说在希腊数据库圈里您很爱讲笑话，能讲个笑话来结束我们的采访吗？

克里斯托·法拉特：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·温丝特：非常感谢。

克里斯托·法拉特：谢谢你的访问。

（张啸剑 译，孟小峰 审校）