

第15篇

拉古·罗摩克里希访谈录

Interview with Raghun Ramakrishnan





拉古·罗摩克里希简介

拉古·罗摩克里希 (Raghu Ramakrishnan) 于 1987 年在得克萨斯大学奥斯汀分校获得博士学位, 1983 年在伊利诺伊理工大学获得学士学位。罗摩克里希于 1987 年加入威斯康星大学麦迪逊分校的数据库系统研究组, 成为该校的计算机系教授。2006 年, 罗摩克里希离开学校, 加入雅虎公司担任雅虎实验室搜索和云计算领域首席科学家。2012 年 3 月, 他离开雅虎公司, 加盟微软公司, 成为其战略技术专家, 领导云计算和信息服务实验室。罗摩克里希和他的研究团队研究了针对聚类、决策树生成、项集计数等多种技术的可扩展算法, 他的研究团队也是最早开展流数据挖掘的研究团队之一。罗摩克里希在 2008 年获得了 ACM SIGMOD 创新奖, 1999 年获得了 ACM SIGMOD 贡献奖, 2001 年入选 ACM 会士。他是 ACM Transactions on Database Systems 副主编, 曾任 Journal of Data Mining and Knowledge Discovery 主编。与人合著的《Database Management Systems》一书是国际著名高校使用率最高的教材之一。他在数据库系统尤其是数据挖掘、查询优化和 Web 数据管理等领域的工作, 对商业数据库系统的查询优化和 SQL:1999 的窗口函数设计都有重要影响。

本专访介绍了演绎数据库，可扩展性，如何轻易烧掉2000万美元？为什么我们要跟决策者接触？

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，今天我来到伊利诺伊大学厄巴纳·尚佩恩分校计算机系采访拉古·罗摩克里希（Raghu Ramakrishnan）。拉古是威斯康星大学麦迪逊分校计算机系教授，他目前的研究方向是数据检索、集成、分析以及挖掘。同时，他是 QUIQ 公司的创始人，该公司开发了一个客户协同的支持系统。他曾经获得美国国家自然科学基金会青年研究者奖、Packard 研究基金奖以及建立和维护 DBWorld 的 SIGMOD 贡献奖。拉古是 ACM 会士，还是非常畅销的一本数据库管理系统教科书的合著者。目前，他还担任了 ACM SIGMOD 主席一职。拉古在得克萨斯大学奥斯汀分校获得博士学位，欢迎拉古！

拉古，首先，我想问一个题外但又非常重要的问题，您的教科书《Database Management Systems》的封面表达了什么含义？

拉古·罗摩克里希：它描述了查询处理的过程。来自威斯康星的奶牛表示用户，其中奶牛的数目和书的版本号是一致的 [威斯康星以奶酪闻名]。问号代表查询，当你提出查询时，箭头会告诉你查询如何执行。你先到 B 树的根结点，然后走到叶子结点，在那里你会找到主键，利用获得的主键就能从关系数据库中找到需要的记录。然后，把这条记录放到缓冲区中，在缓冲区

里使用关系操作对它进行处理。简而言之，这就是关系数据库系统：在内存的层次上计算，用一种灵活的方法移动数据。

玛丽安·温丝特：您从一开始就在“珊瑚”（Coral）项目中研究演绎数据库。演绎数据库已经过时很多年了，但是您在 SIGCOMM2005 发表了一篇名为“声明路由：声明查询中可扩展的路由”的论文，这是不是表明演绎数据库的研究再次复兴？

拉古·罗摩克里希：这篇论文的产生，是源于 Joe Hellerstein 邀请我加入这项工作。Joe 和在伯克利以及英特尔的电报小组一直在深入研究如何在网络路由器中使用回归规则，做了很多很好的工作。Joe 最喜欢说的话是，也许甲骨文或者其它数据库公司会支持演绎数据库，或者思科很快就会推出递归查询处理的产品了。

另外，Serge Abiteboul 和他的同事近期发表了一篇文章，论文中提到了递归的应用。他们提出的系统包含一些信任政策语言的元素，正如你刚才所说的，本质上是执行分布式规则。这篇论文的本质是在高度分布的方式下模拟一个确定的 Petri 网有限状态机的工作。

当然，我们昨天讨论过的信任机制也是使用 Datalog 来工作的。以前我在微软时，有人向我要“珊瑚”的副本，因为他们想在调试时利用递归规则。这些人的工作包括程序的图形化展示等内容，在他们的工作中想借鉴开源的递归语言。

这也充分说明了目前递归规则确实有很广泛的应用。这些应用足以支持工业界吗？我现在还不清楚。但是似乎关注度有所上升。也许是因为在互联网上的节点都可以用递归规则自然地展示出来，也许是因为它的语言类似于 XML 语言（XPath 有递归），也许是因为在安全策略和访问控制中有递归现象。世界上有很多递归应用，我认为我们最终会走到很多递归应用中去。

玛丽安·温丝特：我们一直在讨论旧貌换新颜的东西，那么我想问一个

关于数据立方体的问题，它们是否再次流行起来了呢？

拉古·罗摩克里希：我很高兴看到它们又流行起来了。在数据挖掘领域，传统上，数据库人给关系表带来了可扩展性。对于一个在小数据集上运行得很好的算法，数据库领域的人想要解决的问题是，如何才能使这个算法在大型数据集上也能运行良好。数据库人为聚类、分类、频繁项集等研究内容都做了这项工作。我认为数据库人能给关系表带来比可扩展性更多的内容。我们理解组成的概念，我们知道如何把空间组织用以分析。这是多维数据模型和数据立方体最关注的问题。如何把这些概念和预测模型结合起来或者和其它借鉴统计学和机器学习的工具结合起来？在机器学习的例子中，一个早期关注的标准研究是如何采用一个给定的“案例”并分析这个案例：我该给这个人发放贷款吗？在探索性分析的例子中，强调的重点不在于一个人的案例，而在于如何把数据集看成一个整体来理解：在世界上的哪个地区哪个时间段内贷款决策会有误差呢？为了回答这类问题，可以利用我们数据库界开发的探索性分析工具，并且把它同统计学界开发出来的采样以及预测模型很好地结合起来。我认为这样能产生一些混合工具和混合的方法，从概念上来讲，它们应该更加强大。换句话说，我们不能只是对现有的相关领域的工具进行扩展，而是要开发出新的概念上更广、完全不同类的混合工具，来分析我们遇到的越来越多的数据。

玛丽安·温丝特：在我看来，您的研究兴趣好像走了一圈——又回到了您最初的起点。您现在是完成了所有工作还是要重新开始一个新的周期呢？

拉古·罗摩克里希：但愿不是同一个周期。

玛丽安·温丝特：螺旋上升的周期？

拉古·罗摩克里希：但愿不是一个向下的螺旋吧！我喜欢在一个圈里来回转，我想，也许我会在这个周期多做一段时间。

玛丽安·温丝特：1989年在加州 Laguna Beach 的研讨会上的数据库研究方向报告中，对演绎数据库的未来研究抱有悲观的态度。那么您认为这种报告对演绎数据的研究会产生怎样的影响？

拉古·罗摩克里希：我必须承认，那时我听了报告之后，很不开心。过了这么多年之后，我的反应也成熟了。我认为在一个研究领域中资深研究者的价值就是给后来的研究者一些建议，让他们选择他们想研究什么。但是，我对这类报告的感受还是很复杂的。总体上来说，我希望看到大家关注一些人们认为有前途的研究方向，而不是去关注一些人们认为不应该关注的东西。

玛丽安·温丝特：为什么会这样呢？我看到双方都在发挥作用，做什么以及不做什么。

拉古·罗摩克里希：如果你认为有些事情非常重要又值得你花费时间，你为此耗费了大量的时间和精力，我认为你对这个选择已经考虑得足够多了。现在，如果你认为有些东西并不是那么重要，你主动地放弃了，并轻易地在你认为重要的事情上开始工作。主动地放弃一个别人认为很重要的研究内容，我认为你有可能轻视了这个研究，并没有像一些研究者那样用特殊的视角进行深入思考。所以，我宁愿看到你将自己所有的精力放在一个你深思熟虑过的领域，因为这个领域你洞察得最深，并且最有热情，在这个领域你将产生自己的影响。

玛丽安·温丝特：数据库厂商是如何利用多年来演绎数据库中有关递归规则提升 SQL3 查询性能的研究工作？

拉古·罗摩克里希：我想至少有两个领域。首先关于魔法集重写的想法可以处理相关查询的问题。当查询中有相关的嵌套时，魔法集重写对于面向集的相关性来说是非常有用的工具。我相信有一些厂商已经利用这个技术提高了在 TPC-D 基准测试的性能。第二，在 90 年代，增量视图维护成为一个非常热门的领域。实际上，每个厂商都支持某个版本的增量视图维护，大部

分技术都来源于半朴素评估。这两种核心技术：魔法集重写和半朴素评估起初都是在演绎数据库界发展起来的。

玛丽安·温丝特：在您的职业生涯中，您从纯理论的研究到系统的研究，从学术界到工业界然后又回到学术界，是什么让您做出这些选择呢？

拉古·罗摩克里希：我一直对一些有实际应用价值的研究课题感兴趣。即使在最开始的时候，我的工作是关于递归查询，也是实际应用系统中的一部分。当我是名研究生的时候，我的目标是 LDL（基于逻辑的数据语言）系统，后来我开展自己的研究工作——“珊瑚”系统。因此，我提出的理论几乎都与我参与的一些实际应用工作相关。

这么多年来，我更多的工作是创建和构造系统。事实上，我并不认为这是一种改变。只是我工作比例上的调整，系统和理论的研究都是我一直做的一部分。

玛丽安·温丝特：您怎么看待您的公司 QUIQ？

拉古·罗摩克里希：这家公司是我很难拒绝的一个选择。这是真正做学术圈外的事情的一个机会，可以真正地去做一个若干人使用的产品。这也是一个和毕业于斯坦福 MBA 的兄弟共事的机会，也是一个商业人士和技术人士合作的机会。

经济繁荣时期，钱的问题很容易解决。在几年内，我们筹集了 2000 万美元。创建一家公司是非常有意思的工作，我们做了一些与“雅虎知道”相类似的很酷的事情。从 1999 年开始，我们在为 Ask Jeeves 公司工作，我们用上百万个不同的警报来连续监控 Ask Jeeves。我们的公司有像 Compaq、Sun、National Instruments 这样的客户。我们试着在技术支持中使用集体同作及答疑的理念，但是最终的结果却不尽人意。

玛丽安·温丝特：告诉我们结果吧！

拉古·罗摩克里希：我们按照自己的工作方式花掉了 2000 万美元，比预计得要快一些。我们用了更多时间来实现技术到商业策略的转变。让人难过的是我们最终没有想出解决方案。Compaq 公司的终端一直广泛使用我们的技术，当然他们会付给我们很多钱，他们对所获得回报感到非常满意。有几家大公司已经成为了我们的客户，但是就在我们试着取消一些生意的时候，由于入不敷出，我们的公司陷入了危机。再加上经济下滑，我们没能尽快地取消那些生意。此外，还发生了几件事情，我们与苹果公司的一个关键生意失败了，因为与我们谈生意的人在我们开始为他们工作的时候被解雇了。在这种大环境下，风险投资人并不想做投资。就如我所说的，我们不得不在对公司都很不满意的时候卖掉公司，其实我们在更早的时机可以采用更优雅并获利更多的方式退出这个舞台。

然而，建立一家公司真的是非常令人兴奋，就算在公司每况愈下的时候，我们依然勒紧裤带带着员工前进。这也是我觉得做得最困难的几件事。

总的来说，这是一个我非常珍贵的独特经历。

玛丽安·温丝特：您是如何再回到学术圈的？

拉古·罗摩克里希：非常困难。我认为风险投资家会喜欢我最初的一些建议，但是我的同伴并不这么认为。我的整个价值体系扭曲了。我过去一直寻找一些在商业上可能、但是在学术上并没有开发出有新意的东西。我花了一段时间去调整这种意识。

玛丽安·温丝特：问题反过来想一起，NSF 会喜欢您的建议吗？

拉古·罗摩克里希：我告诉过你，我们很容易地用掉 2000 万美元，没有挣到太多的钱，剩下的你可以自己得出结论。

玛丽安·温丝特：最初，数据挖掘的研究依托于数据库领域，但是现在它好像已经发展成一个独立的领域，而且有三个独立的会议。这两个领域的关系是怎么样？

拉古·罗摩克里希：我觉得这两个领域之间像现在这样的互为补充的关系非常好。数据挖掘相对于数据库来讲包括更多的内容，机器学习和统计学都是数据挖掘中不可缺少的组成部分。对于数据挖掘领域来讲，它有自己的定位，试着形成一套只属于数据挖掘的基础理念是很重要的，这些理念也是数据挖掘和其它领域的交叉点和区别。

同时，我非常希望我们一直能看到数据挖掘相关的论文发表在主要的数据库会议中，也希望看到数据库技术的研究者参与数据挖掘会议，我们不想看到这两个领域分裂。显然，这样是对数据挖掘领域有益的；我认为，在将来数据库领域将会更多地关注怎么看待数据的问题。现在很多有关数据挖掘方面的问题都与这个目标非常相关。例如，如果你正在做系统调优方面的工作，可以借助机器学习技术来帮助调优。会有哪些其它的技术对你的场景有帮助？谁是你可能合作的人？我想我们可以运用这种协同优势。

玛丽安·温丝特：最近您开始关注数据隐私和安全领域。这种研究引入了很多数据库研究者没有考虑过的方面：法律、道德和公共政策。比如，如果我发布了一个连接函数比原来的方法快了 20%，没人真正关心我是不是错了。但是，如果我说用新的数据匿名技术可以比原来的方法保护数据隐私度高 20%，那就是一个吓人的承诺：公众就会认为这意味着什么。您是不是要抓住这些新方向，您打算如何掌控它呢？

拉古·罗摩克里希：这是一个非常好的问题。我最近花时间和两个人一起研究数据挖掘和法律的文章，他们是加州大学伯克利分校的法学教授 Deirdre Mulligan 和普度大学的 Chris Clifton。这是一个真正学习的过程。我们这些技术研究者正在开发一些能够普遍应用的工具。有些言论认为我们不该开发这些工具，我认为这是错误的。如果我们不做，还会有其他人来开发这些工具。同时，我不认为恰当的反应是“这是技术问题，我所做的一切事情就是开发它，我对其他人会对它做什么概不负责。”

玛丽安·温丝特：您是否听说过前几年政府给所有的 Google 搜索发了法院传票？

拉古·罗摩克里希：是的，Google 似乎总是在触犯法律的边缘做开发。

玛丽安·温丝特：但是和我们一样，他们的问题也是规模上的问题。

拉古·罗摩克里希：他们是规模的问题，也有版权的问题。想一下 web 上的数据库问题：所有权、法律权利、隐私——这些都会涉及到，且不可分割。所有权是这些问题的核心。无论你是在讨论版权的所有权，或者是隐私信息的所有权，这些东西是网络上的一个灰色地带。

回到你最初的问题，我认为数据库研究者应该考虑如何将不同领域的工具拓展，使其能够为数据隐私和安全中与其它不同的设置的折衷服务。这个意思不是我们去设置政策，但是我们可以给制定政策的人一些指导。我们必须为制定灵活的政策提供工具上的支持。

从算法方面或者理论证明方面来说，我们应该考虑下面的事情：不要只是证明算法到底有多快，要证明你的定理能走多远，比如，破解某种巧妙的匿名方法。这件事并不单单是在研究上是明智的，我还认为能做一些传统上没有做过的事情是我们的责任。社会领域和立法领域都需要我们去影响制定政策的人。我们应该告诉这些人，因为他们也许不知道现在收集数据以及对数据的分析带来的风险。也许我们不清楚将所有数据都收集起来会发生什么，但是我们肯定知道什么事情一定会发生的概率。我们清楚地知道哪些工具可用在这里，以及简单地开发出新的工具。我觉得我们应该在社会事务中更加积极一些。

玛丽安·温丝特：从传统上说，我们不善于完成这类事情，它是怎样发生的，谁来做这件事呢？

拉古·罗摩克里希：是我们，尽管我们之前没有做过这些事。

玛丽安·温丝特：我的意思是“我们”指谁呢？

拉古·罗摩克里希：你和我。在大学里任职的人们，工业界的领军人物。我们，不是指有近期研究压力的年轻人。像我们一样年长的人有时间、有远见，不局限于现在就能做的研究，而是关注于未来可以做的研究。我们，应该是做这些事的人。

玛丽安·温丝特：您会怎样结束数据库研究？

拉古·罗摩克里希：这是一个非常有趣的问题。我到得克萨斯大学学了 Hank Korth 的数据库课程，我真的非常喜欢这门课。我不知道我如何敲开了 Avi Silberschatz 的门，但我确实做了，然后我们开始在一起工作。我的第一篇论文是关于并发逻辑程序的，发表在 POPL (Principle Of Programming Languages) 上，我认为那不是我的理论领域。然后 Avi 指导我去做一些数据模型的工作。然后我就到了 MCC 并且在一个完全新的领域——演绎数据库中找到了自己的位置。我有幸和几个非常出色的人一起工作，他们非常优秀，像 Catriel Beeri、Francois Bancilhon、Carlo Zaniolo 等人。我正好有这个背景，我懂逻辑程序也懂数据库。所以我跳槽到那儿，并且一直工作到现在。

玛丽安·温丝特：当您在奥斯汀做研究生时如何安排时间？

拉古·罗摩克里希：这就要回到那个具有共享大型机的时代。白天的工作效率非常低，停车场也很拥挤。所以，我通常晚饭之后开始工作，早饭之后开始睡觉。这样的话，停车就很方便了，机器也像现在的机器一样快了。有些人发现我有斜视症，那是因为我原来晚饭只吃麦片粥。最后，Bancilhon 回到法国，他的离别感言是“我终于和拉古的作息时间一致了。”

玛丽安·温丝特：您吃过的最辣的食物是什么？

拉古·罗摩克里希：我曾经有一次误入了一家泰国餐厅，点了餐，并让他们做得辣一点。虽然我喜欢吃辣的食物，但是我再也不会去泰国餐厅点辣

的食物了。

玛丽安·温丝特：您能跟我们描述一下它有多辣吗？

拉古·罗摩克里希：辣得连回忆起来都觉得痛苦。

玛丽安·温丝特：过去的二十年对于专业团体来说是很困难的，会员数量有很大程度的减少。计算机科学的团体也没能避免。您认为 ACM SIGMOD 在数据库研究者的生涯中能起到什么作用？

拉古·罗摩克里希：首先，它能提供机会，尤其是对于信息技术、数据库。像印度和中国这样的国家，专业人员在大幅度增加，但是来自那些国家的注册的会员数却少得令人难以置信。我们需要尽量使我们的会员国际化，并且能在一年之后保证他们还在组织中。

玛丽安·温丝特：对于印度和中国的会员，您能提供什么样的政策让他们愿意加入呢？

拉古·罗摩克里希：我们正在完善网络和通信。大家从这里得到的好处是：我们是一个整体。

玛丽安·温丝特：他们不能从网上得到这些吗？

拉古·罗摩克里希：我认为网络是一种机会。过去，会员最大的好处是，你可以亲自参加一个多数是在美国偶尔在欧洲召开的学术会议并且任职。网络的好处在于：我们可以把会议现场做成视频，放到网上，大家可以看到这个视频，这需要钱。一部分钱来自于会员费。这是一个商业模式，因为 ACM SIGMOD 到最后就是一个生意。我们不关注于是否盈利，只要不亏本就可以了，我们也不给任何人发津贴。我们需要告诉大家它是如何运作的，比如会议、专题报告以及其它活动，并且能让我们领域的其他人参与，不管他们在哪里。这就是 SIGMOD 会员可以享受到的便利，幸好他们都认为这值得让他

们付钱，这促使我们给更多人提供便利。

我们也需要在世界上较偏远的地方开会。我们正在做这件事情。SIGMOD 已经做了有意识的努力，走出去，走到每个地方。我希望看到这种努力可以继续下去。我认为像 SIGMOD 这样的组织能发挥领导作用，让 SIGMOD 以外的、同时与 SIGMOD 具有相同兴趣的组织（比如 SIGIR、SIGKDD 以及其他的 SIG 组织）聚集到一起。也许我们有这种创造力，能和其他的 SIG 组织一起工作，创造互相帮助的机会。

玛丽安·温丝特：您以前发表的论文中，有您最满意的工作吗？

拉古·罗摩克里希：有两个我最喜欢的工作，第一个是我写的魔术集论文中的一篇，我花了 6 个月的时间试着弄清楚怎样去概括 Bancilhon、Sagiv、Maier 和 Ullman 写的各种版本，但是解决的方法在五分钟的咖啡时间就找到了，我一直记得当时我叫了一声“啊哈”。另一个最满意的工作是在 SSDBM 中处理关系排序的工作，这个工作是鲜为人知的，但这个工作最终对 SQL99 的窗函数产生了很大的影响。我一直对流和序列有很浓厚的兴趣，这是很有影响的工作，也是我喜欢它的原因。

玛丽安·温丝特：如果您有很多时间在工作中去做一件您应该做但现在还没有做的事情，这件事情会是什么？

拉古·罗摩克里希：我可能会经常和我的同事共进午餐。

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

拉古·罗摩克里希：我想吃点能让自己变聪明的药片。

玛丽安·温丝特：大家都说您非常幽默，您有什么笑话和我们分享吗？

拉古·罗摩克里希：我告诉你一个关于比尔·盖茨和印度企业家的故

事。很早以前，一个印度企业家访问雷德蒙德，告诉比尔：“你应该到印度去招聘员工，因为那里已经开始行动了”。但比尔并不感兴趣。他说：“不，跟我来”。他给了这个印度人一个铁锹让他挖。这个企业家挖呀挖，遇到一根电缆，这时候比尔说：“有线，这就是要干的事。”（这就是比尔准备购买有线电视公司的期间）。这个印度企业家非常失望地回去了。最终，比尔看到机会，到印度去招聘员工了。（微软是最早直接到印度去招聘员工的美国公司之一。）比尔·盖茨的印度企业家哥们给比尔一把铁锹，说：“挖。”比尔欣然地挖了起来。那时候印度非常热，五分钟过后，比尔出汗了，十分钟，十五分钟之后，他膝盖弯曲，毕竟他已经不年轻了，所以他感到非常疲惫。他抬起头，这个印度人告诉他：“无线，这就是要干的事”。

玛丽安·温丝特：谢谢您与我们谈话。

拉古·罗摩克里希：别客气，这是我的荣幸。

（霍峥 译，孟小峰 审校）