

第12篇

珍妮弗·威多姆访谈录

Interview with Jennifer Widom





珍妮弗·威多姆简介

珍妮弗·威多姆 (Jennifer Widom) 是斯坦福大学计算机系 Fletcher Jones 讲座教授，曾任计算机系主任。珍妮佛·威多姆于 1982 年在印第安纳大学音乐学院取得学士学位，并分别于 1985 年和 1987 年取得康奈尔大学的计算机硕士和博士学位。博士毕业之后她进入 IBM Almaden 研究中心工作至 1993 年，随后进入斯坦福大学任职至今。她的主要研究领域涉及数据管理的多个方面，包括 XML 数据库、数据流、不确定数据管理、众包等。珍妮佛·威多姆是 ACM 会士，同时也是美国国家工程院及美国艺术与科学院的院士。她与 2007 年获得 ACM SIGMOD Edgar F. Codd 创新奖。同时，她还是一系列重要会议的程序委员会成员。

本专访介绍了运气，成功要素，开辟新领域的时机，选择好丈夫的重要性，惊悚的假期，成为助理教授有多难等。



玛丽安·温丝特

玛丽安·温丝特：欢迎来到 ACM SIGMOD Record 数据库领域杰出人物访谈。我是玛丽安·温丝莱，现在我们在伊利诺伊大学厄巴纳-香槟分校的计算机科学系。坐在我旁边的是珍妮弗·威多姆，她是斯坦福大学计算机科学与电子工程系教授。珍妮弗·威多姆目前主要的研究领域包括：数据溯源、不确定数据管理、互联网数据查询和数据流。珍妮弗是 ACM 会士、Guggenheim 会士和美国国家工程院院士，在加入斯坦福大学之前，她任职于 IBM Almaden 研究中心。她在美国康奈尔大学获得博士学位，现在，欢迎珍妮弗的到来。

珍妮弗·威多姆：谢谢！

玛丽安·温丝特：珍妮弗，您把团队名称由“The Database Group”改为“The InfoLab”是不是意味着数据库研究已经走向消亡呢？

珍妮弗·威多姆：不不，当然不是。其实我本人是做数据库核心技术研究的。我的大多数学生也是如此，并且会持续下去。不过，我也有个别学生确实在做数据库领域之外的研究。之所以要改团队名称，主要是为了更好地涵盖所有学生的研究领域。比如，我有一个学生在做图片检索和标注方面的博

士论文。我们不想让每个学生都被冠以数据库研究的标签。我们有的学生在做信息检索的工作，我也不想让他们被归类为数据库研究者。数据库研究是 InfoLab 的研究领域之一。

玛丽安·温丝特：您似乎有超前的能力来选择有前途的研究领域并成为该领域的领军人物。在您的研究过程中您是如何做到这一点的？

珍妮弗·威多姆：这中间或多或少都有一些偶然性。最初我们在做 LORE（一个半结构化数据库管理系统）项目时，它只是我们在使用半结构化数据模型做数据集成项目的一个小分支。后来，我想我们为什么不能创建一个提供半结构数据存储和查询的系统呢？我们确实这样做了，经过我们对项目的不断完善，最终诞生了 XML 数据库系统。

数据流研究也是这样，我关注数据流研究已经有很长一段时间；我认为这个方向应该非常有趣，不过我无法说服我的学生从事这方面的研究。最后，我有几个对这方面感兴趣的学生，我们推出了一个项目。此后，数据流领域的研究开始兴起。

至于我正在从事的研究领域，我一直认为需要开辟一个新的研究方向，但是我确实无法确定应该从事哪方面的研究。每天晨跑的时候我都会思考如何能够在大量的应用场景中将数据不确定性和数据溯源研究结合起来。我考虑结合数据自身建立一个系统来处理这两个方面的数据。

所以我想说所有这些研究方向都是相当随机的选择，它们几乎都是在下意识的情况下决定的，没有一个重大的思维过程或宏大的视角。我不认为自己是一个有远见的人。

玛丽安·温丝特：您如何判断什么时候离开当前的研究领域并进入一个新的领域？

珍妮弗·威多姆：在某种意义上，我可能会比较早的离开当前研究领域。即便原有研究领域仍有大量的工作需要去做我也会坚定地改变研究方向进入

一个新的研究领域。促使我做出这样决定的因素主要有两个。首先是学生的研究兴趣。假如我们已经在项目工作好几年、并构建了相当成熟的研究原型，那么，即使我有五个清晰的研究主题，如果对我的一年级博士生推荐这方面的工作，他们不会对其中的任何一个感兴趣。对于一个研究了 4 ~ 5 年的方向，学生们不想在这样的“旧领域”继续对已经定义好的研究框架继续展开研究。因此，我的学生们差不多每隔 5 年的时间就会开辟新的研究领域。

其次，如果一个研究领域已经吸引了许多研究者，我会倾向于将研究点转向少数人在研究的新领域。我喜欢尽早参与，也会选择尽快离开。有时我会用冲浪来类比我的研究工作：你正骑在一个浪尖，然后在适当的时候退出，让波浪继续。我喜欢这样做。

玛丽安·温丝特：对于如何在一个新的领域开始新的研究项目，您是否能给我们一些建议？

珍妮弗·威多姆：在开始一个新的项目时，我建议你花一年左右的时间了解基本知识并勾勒出如何将研究工作适用于新领域的系统框架。如果你正在从事一个新的研究点，那么必须要花费很长的时间来学习基础知识，我认为基础知识是非常重要的。不过当遇到非常实际的应用时，你必须要意识到是时候来进一步推进工作并开始构建应用系统。结合获得的应用需求和数据进行研究。

玛丽安·温丝特：能说说目前为止您认为自己最成功的研究项目吗？

珍妮弗·威多姆：我判定一个项目成功的标准是看工业界人们是否对我的项目成果感兴趣并开始构建类似的系统和应用。现在我觉得数据流项目一直是最成功的。工业界对这个领域已经产生了浓厚的兴趣。人们被我们项目取得的成果吸引，这其中就包括我们开发的查询语言。

其他项目中我感觉最成功的是 LORE，我们构建的系统最终被用于 XML。这个项目因为 XML 而变得非常出名，而且我认为它遇到了一个幸运

的时机。我们构建一个基于半结构化数据模型的数据库系统, XML 非常接近这个模型。我们没花多少精力就把我们的模型转化成 XML, 我们突然就在很早的时候拥有了 XML 系统。因此, 我认为 LORE 是一个非常成功的项目, 这其中也有一些非常幸运的时机。数据流项目就没有这样幸运的时机。

玛丽安·温丝特: 您是否认为工业界是从学术界哪里获取灵感呢?

珍妮弗·威多姆: 不, 我不这样认为。我认为工业界产生了业务需求, 然后开始探求用于解决这些需求的技术。据此, 我发现了数据流的存在: 来自不同领域需求的人们谈论他们需要的技术, 这些技术看起来像数据流, 像连续查询。他们在验证了什么是他们需要的技术后, 发现了流处理技术。

玛丽安·温丝特: 对于大卫·德威特 (David DeWitt) 提出的流数据管理不需要专门的数据库系统这一观点, 您是怎么看的?

珍妮弗·威多姆: 这或许是正确的, 我不打算争辩说, 我们绝对必须有一个专用的数据流系统。我们从零开始构建了数据流管理系统, 这其中我们享受了很大的乐趣, 但是, 是否可以说它将是数据流未来的主流研究方式? 不一定。

有些公司构建了跟我的流处理基本相似的系统, 也是一个原生数据流处理系统。有些公司的产品是基于数据流而连续查询隐藏在软件中。我相信主要 DBMS 供应商正在将流技术添加到它们的系统中。数据库的其他技术也是沿着这个方向发展起来的。起初, 小的构建原生系统的创业公司开始流行。XML 数据库和面向对象数据库系统就是这样兴起的。我相信还有很多相似的案例。此后大公司会说: “嘿, 我们可以添加到我们的系统。这不是那么难, 我们已经有这方面的技术, 它们之间可以做到无缝对接。”接着, 这些小公司就会消亡。因此, 我不打算争论关于流数据管理不需要专门系统这一说法。

另一个不同的说法是我们可能不需要数据流技术。对此, 我并不赞同。我认为在过去的一两年, 工业界已经表明需要以不同的方式寻找流数据, 需要

连续查询。

玛丽安·温丝特：您认为目前从工业界涌现的杀手级应用有哪些？

珍妮弗·威多姆：目前，金融监管是主要的应用之一。另一种是网络点击数据流，通过这些数据流可以实时地获知人们在浏览哪些网站以及在这些网站上所做的操作。这种应用我们称之为业务行为监控。利用这些应用，人们试图通过数据流的方式记录在他们业务运营过程中所发生一切，这些数据就像指示板那样，可以查看并允许他们的高级管理层做出商业决定。这里我们列举了需要数据流处理技术的三种应用场景。

玛丽安·温丝特：那么您认为您研究的领域会朝什么方向发展呢？

珍妮弗·威多姆：我可以吹响我自己的号角，并告诉你关于我的下一个项目，三重奏。我对不确定数据管理非常感兴趣，我认为人们需要 DBMS 越来越多支持不确定数据的相关操作。

人们对不确定数据的研究已经持续了很长时间，但这方面的研究并没有成为主流。现在我们重新来看的话，人们愿意把他们的数据的确定性信息存入数据库，并开始查询。但上述情况事实上确实还没有发生，因为当你与人交谈，他们往往告诉你他们的数据并不是非黑即白。这是我对不确定数据感兴趣的原因。

这个话题似乎与谱系研究颇为相似，因为如果你的数据是不确定的，你可能会想知道这些数据是从哪里来的，他们是如何演变的，你会试图找到有质量的数据。数据质量是很重要的。

就其他领域来说，对数据集成的研究会持续进行，人们在该领域还有很多工作要做。因此，数据集成在数据库领域会占据很大一部分领地。究竟人们会怎样开展研究工作，我不太确定。我的一些学生目前在关注网络服务并且尝试对错综复杂的网络服务进行查询集成。

数据集正变得越来越庞大，越来越繁杂。数据清洗工作是很重要的。

玛丽安·温丝特：珍妮弗，您在硅谷的核心区域几乎渡过了您的全部职业生涯，可您从未进行创业，这是为什么？

珍妮弗·威多姆：我不是一个创业类型的人。我可以简单解释一下。我们在做 LORE 项目时确实想过进行创业。但我发现我们需要找到投资人并且我自己也不具备做 CEO 的潜质。对于这样的设想起初感觉很有趣，但后来我失去了兴趣。我喜欢自由支配自己的时间，自己做老板。即使你是一个公司的老板，总有别人实际上是你的老板，你的老板可能是你的投资者或你的客户。我不热衷于这个。对事物我想有预见性，而创业是非常难预测的。对于我来说，创业不会给我带来快乐。因此，我一直都非常乐意在顾问委员会为他人提供咨询。对于我来说，这是和硅谷进行沟通的一个非常好的方式。通过做一些咨询和与人交谈会让我觉得我做的研究与实际应用是紧密相关的，这样我也不需要处理创业所要面对的诸多繁杂的事物。

玛丽安·温丝特：您是我采访的有孩子在上小学的几个人之一，您是怎样做到既在研究领域获得丰硕成果，又能够很好教育两个孩子的？

珍妮弗·威多姆：我认为要很好地既能在事业上有所成就也能教育好子女的关键因素在于拥有一个很好的丈夫。我的先生也是一位教授，我们能够非常和谐的共同处理遇到的每一件事，这样抚养孩子并没有花费我太多的精力。所以说，对于那些观看或阅读这篇采访的人来说，（寻找伴侣时）一定好好思考以免为时过晚。

找到一个合适的丈夫非常重要，接下来最重要的是高效地开展工作，要知道哪些事情是重要的，哪些是不重要的。也许通过忽略一些不重要的事情，你可以把时间和精力用在真正重要的工作，这样你就能腾出时间和家人在一起。

玛丽安·温丝特：您每天晚上睡几个小时？

珍妮弗·威多姆：有时候我不能保证有充足的睡眠。忙的时候我只有

5 ~ 6 个小时的睡眠时间。对我来说，每天有这些额外的几个小时的时间来管理家庭和工作上的事情是绝对重要的。

玛丽安·温丝特：您是怎样让您的办公桌显得如此整洁？

珍妮弗·威多姆：到过我办公室的人都知道我的办公桌是非常整洁的，我真的很喜欢把那些不用的东西丢掉。我是一个“丢弃主义者”。许多人都习惯于将多余的东西打包整理，我正好相反。我有一个理念，如果你抛开一切，或者几乎一切，那么，你会发现花费在重新获得你扔掉的那些事情的时间远低于你用在处理那些没有扔掉的事情上的时间。换句话说，花费在处理不小心扔了一些重要的东西上的时间通常是相当低的。花费在处理所有你认为可能是重要的东西上的时间是相当高的。我会扔掉办公室里的大部分东西，甚至是计算机。这样我的办公室才显得非常整洁。

在去年的 SIGMOD 会议上，我有一个由微软戈登·贝尔做的题为“我的生活片段，一个用于处理个人事务的数据库系统”的报告。（相关链接的网址是 <http://research.microsoft.com/barc/mediapresence/MyLifeBits.aspx>）。那个报告讲述的是我们应该完全记录每个人生活中遇到的事情，保存这些生活记录并使它们成为可用信息。举个例子，我们可以记录昨天通话的电话号码，这样以防有一天再次用到这个号码。我完全不想这样做。尽管保存这些信息可以很快地得到电话号码，因为查找这些信息需要花费一些时间，但我认为这不是一个利用时间的好方法，与我的理念是完全违背的。我会在早些时候把这些信息按效率进行分类：尽量更加高效的生活，这样才能有时间去工作和照顾家人。

玛丽安·温丝特：您的大学专业是音乐。为何取得音乐学位后却最终去做数据库研究？

珍妮弗·威多姆：我在印第安纳大学音乐学院攻读小号表演专业，我们有一个很特别的要求，就是至少选修 3 门同音乐理论和表演无关的课程。其

中一门我选修的也是本院开设的，称之为“计算机在音乐研究中的应用”。我只是很随意地选择了它作为我的选修课程。我们编写 SNOBOL 程序来分析表示音乐的流，我被这个吸引住了。那时我大学三年级，我开始去学习印第安纳大学开设的一些计算机课程。我完成了小号表演专业的学业，然后留在了印第安纳大学。我继续从事音乐，但是我已经转入攻读计算机专业硕士学位。问题是我没有计算机本科的学习经历，怎么能做到这一点。当时学校就是在我仅有选修几门计算机课程的基础上同意我可以主修计算机专业，之后我确实不断拓展自己，也参加一些研究。在我完成了这种“学士学位”，实际上是硕士学位之后，我决定去康奈尔大学攻读博士学位。这就是我从音乐转换到计算机科学的经历，我一直演奏小号到 1992 年，那时我已经完成博士学业一段时间了。

玛丽安·温丝特：什么使您放弃了小号？

珍妮弗·威多姆：我厌倦了练习。这有点像运动，尽管有些人不这样认为。在 IBM Almaden 研究中心工作那段时间我每天练习一个半小时，而且还活跃于圣何塞附近的音乐圈。有一天我意识到其实我根本不想再重复这种练习。我不想这样反反复复，而演奏小号就是这样的体力活动。因此我决定放弃。不过我现在正在考虑重操旧业，因为我儿子正在学习小号，他需要人进行二重奏。

玛丽安·温丝特：您在工业界实验室和大学都待过，您如何看待自己在 IBM 的那段日子？

珍妮弗·威多姆：在 IBM 的那段日子很棒。那是很惬意的时光。在 IBM 的时候，我们真正地只需要去做研究。没有很多的管理责任，你不用像教师那样去申请基金。我们真的有很多特许。我在一个从事 Starburst 项目的小组中，主要是构建一个用于实现我们研究想法的优秀基础架构的大原型。我将我的大部分时间都用来做研究；我相信我在 IBM 的那段日子是我研究生涯的

一个高峰。因此我在 IBM 的日子充满诗意。

我关于数据库方面的知识也大多是在 IBM 的时候学习的。我博士方向是编程语言，因此在 IBM 的时候可以很自由地去学习数据库，那里有着当时世界上最好的数据库研究小组之一。我在那里度过的 5 年真的很棒。

玛丽安·温丝特：什么使您从工业界实验室转向学术界？

珍妮弗·威多姆：我是一个教授的孩子；我一直认为成为一个教授是件非常棒的事。在我的骨子里流淌着学术的基因，因此一旦有了去斯坦福大学担任教职的机会我没有丝毫犹豫。

玛丽安·温丝特：您考虑过回到工业界实验室么？

珍妮弗·威多姆：从没。我喜欢当一个教授。这是世界上最棒的工作，我对此深信不疑。

玛丽安·温丝特：编程语言界和我们数据库界有什么不同？

珍妮弗·威多姆：很不同。我取得了编程语言的博士学位，那几年也去了一些相关的会议，我发现数据库界更友善，更社交化，少了自私和言行不一。我不是想贬低编程语言同行，但是我发觉在数据库圈子里真的感觉很轻松。我认为这可能和基金情况有关。现在不论哪个领域都很难获得基金了，但是曾经有段时期数据库领域比其他领域更容易获得资助，同时数据库领域跟工业界有着更强的联系，他们在待人处世时有着更多的自信，而编程语言领域可能不是这样的。相应地，数据库界也有更多热衷于社交聚会的人。

玛丽安·温丝特：您喜欢跟全家去国外度假（当我准备这个采访的时候，我的一些“线人”用“古怪的”和“危险的”等词语来描述您的假期。）当海啸快要来的时候您怎么会跟您的丈夫和孩子在泰国冲浪？

珍妮弗·威多姆：在你的问题中有一些误解。我不会去争辩“古怪的”和

“危险的”这些词——嗯，我们不会将家里人置于危险之中，但也许有些人认为我们的旅行是古怪的。我们的确有去国外的冒险假期。另一个误解是我不认为谁会知道海啸即将到来，对吧？我只是认为它会来。第三个误解是海啸来时我们并不在泰国。我们想在泰国待的时候能租个帆船到处转转。但是我们想要的帆船那个时候没有，所以我们转而去了新西兰。实际上，当导游提醒我们一场大海啸袭击了泰国时我们正在新西兰激流泛舟。我们想，哇，我们本应该在那划船的！但是我们没有，我们很好，第二年我们去了泰国，租了个帆船去被海啸袭击的地方看了看。

玛丽安·温丝特：Jeff Ullman 的公司设计了一个可以自动产生家庭作业的系统，您喜欢用他的 Gradiance 软件吗？

珍妮弗·威多姆：我是 Gradiance 的大“粉丝”。事实上，我跟 Jeff 讨论了这个系统的很多方面，但我真的很喜欢它，所以把这些话写出来。Gradiance 从一个问题库中产生家庭作业问题，并修改作业和给出反馈，这些全是自动的。学生可以反复地练习，每次得到不同的题目，且可以实时得到分数。学生很喜欢它。Gradiance 也有一个 SQL 引擎，主要是为了刚刚接触数据库的学生。你给 Gradiance 一个模式和数据，且用英语给出查询。学生用 SQL 写查询，这些查询就在数据库中运行，提供实时的反馈。Gradiance 会告诉学生他们得到了错误的答案，并给出数据和正确的答案。然后学生可以再次尝试这个查询。Gradiance 有一些非常先进的技术，比如它有一个额外隐藏的数据库，这样学生就无法在他们看到查询答案后再次去愚弄系统。学生对这个 SQL 引擎和它提供的功能反响很好。因此我认为 Gradiance 是一个很优秀的教学工具，我很喜欢用它。

玛丽安·温丝特：我听说您认为做报告的技能是极其重要的，能就这点展开来谈谈么？

珍妮弗·威多姆：就口头报告而言，我们小组的学生需要进行实际的报

告并接受批评，我们小组因此而名声在外。我认为我们小组在如何做一个好的会议报告以及做好的报告的重要性方面有很高的要求。因此我的学生会反复地进行报告直到他们觉得可以了。同时我们也做很多的指导。

就写作技巧而言，我认为也是一样的。我认为写一个清晰的文章极其重要，也是非常困难的。我们花了很多时间去讲什么是构成好文章的要素，什么不是。我对学生的草稿非常挑剔。在我第一次给他们修改完文章时，文章基本会被改得面目全非。

我一个高年级学生自认为文章写得极好。他认为写得太好的文章容易使评阅人挑到毛病，因为他们真正理解了这些文章。因此他决定对他写得很好的一篇文章进行一个实验，就是将它改得不那么好，然后提交到会议去看看是否会得到更少的差评。他坚信结果会如此，不幸的是，我也认为事实或多或少就是这样。

玛丽安·温丝特：有时候评审人不确定您在说什么的时候您会从中获益

珍妮弗·威多姆：没错。我认为这经常发生。一个人评审文章越快，越有可能认为你的文章没做什么好的工作。

玛丽安·温丝特：您职业生涯的下一步是什么？您会设想成为系主任么？

珍妮弗·威多姆：在最初的时候，我可能会认为成为系主任很有意思。现在我可不这么认为了，理由跟我不创业的原因差不多：当你是主任的时候，你开始对重要的事情失去控制。你对你的时间表失去了控制，每时每刻你都得打扮好（我不擅长打扮），而且你还要和不同人打交道以便获得大的捐款，诸如此类的事，我不认为这些是我喜欢的。正如我前面所说，我不是一个有远见的人，而我认为系主任应当具有远见。所以，我不认为系主任的职位对我有什么吸引力。

事实上，我职业生涯的下一步是，我们一家将暂停工作，去进行为期 14 个月的环球旅行。当我回来时，我猜我可能会有一些清晰的想法。

玛丽安·温丝特：您对数据库界的新人、职业生涯中期的研究者和相关的从业人员有什么建议么？

珍妮弗·威多姆：我只能对研究人员给出建议。现在成为一个年轻的研究者已经十分困难，这早就不是什么秘密了。我认为人们不该去回避这个问题。现在的确比以前难了。如果你想成为一个助理教授你必须充满着激情。你需要去获得基金，而这比以前要困难不少。因此这并不容易，你必须真的想得到它。我的建议是保持对工作的热情和激情，永不气馁。

我认为最重要的建议是“永不气馁”。眼光需要放长远。如果你的文章被一个会议拒了，可能并不是因为你是一个很糟糕的研究者。只要等待下一个会议并从长远来看，尝试将你的工作变得具有全局影响力而不只是担心某个特例。

玛丽安·温丝特：听起来您认为主要会议的接受率应该更高点。

珍妮弗·威多姆：也许吧。关于会议、出版以及其中的问题我们可以做进一步的深入讨论。我认为其中存在问题，关于这个过去这些年已经讨论很多了。我不确定什么是最好的解决办法。有些人在探讨没有接受率限制的在线杂志，仅仅依靠质量来衡量，我认为这是个有意思的主意。这是个很复杂的问题，但是我真的认为当前在我们有选择性的会议中，很多有价值的文章没有被接受。正因如此，我很担心年轻人的职业生涯。现在的学生在发表论文上压力更大，如果你想去寻求一个教职的话。现在的情况很紧迫。

玛丽安·温丝特：假如您有足够的时间去额外做一件你现在工作没法做的事情，那会是什么？

珍妮弗·威多姆：我没有足够的时间去研究计算机科学的其他领域，或是计算机之外的其他领域。（我怀疑这是不是这个问题最普遍的回答。）我想对各种事都了解更多。即便是很与数据库关联的事，比如信息检索、数据挖

掘——我是说，这些是我的领域中很具有实践性的，对此我了解得不是很多。人工智能、自然语言理解——这些都是我应该了解更多的。此外还有些仅仅只是想了解的。我想多了解点生物学，因为它十分流行和有趣，还有绘画和其他的所有领域。

有时间去学习这些事情将会非常好，但是我目前看不到希望。也许当我的孩子读大学时我可能有时间。

玛丽安·温丝特：作为计算机科学家如果您可以改变关于您自身的某件事，您希望是什么？

珍妮弗·威多姆：也许我想多做点编程的工作。我基本上是将自己看作一个从事系统研究的人，我的学生也都在构建系统。我想更多地跟那些系统接触，深入了解系统的细节，因为我现在实际上已经脱离系统内核了。我更想知道系统里面是如何运作的，甚至我想参与去构建这个系统。那会很棒，但我现在没有足够的时间。

玛丽安·温丝特：珍妮弗，谢谢您今天接受我的采访。

（慈祥，王江涛 译，孟小峰 审校）