第14篇

黄奎勇访谈录

Interview with Kyu-Young Whang

黄奎勇简介

黄奎勇（Kyu-Young Whang），韩国人，ACM 会士，IEEE 会士以及 IFIP WG 2.6 会员。1973 年以优异成绩毕业于首尔国立大学，1975 年获得韩国高等科学技术学院（KAISI）硕士学位。之后赴美深造，1982 年获得斯坦福大学硕士学位，1984 年获得博士学位。1983 ~ 1991 年期间，他在 IBM T. J. Watson 研究中心任职。1990 年，他回到韩国加入 KAIST，成为 KAIST 计算机科学系著名教授以及先进信息技术研究中心主任。他的研究兴趣十分广泛，包括：数据库系统 / 存储系统，面向对象数据库，多媒体数据库，地理信息系统，数据挖掘 / 数据仓库，XML 数据库以及数据流。他在国际期刊和国际会议发表论文已超过 110 篇，在韩国国内发表论文 170 多篇。作为 VLDB 杂志的创办人之一，他在杂志编委会服务了 13 年之后，在 2003 ~ 2009 年期间成为杂志的主编。曾任韩国信息科学学会（KISS）主席。领导开发了 Odysseus 数据管理系统，实现了 DB 和 IR 的结合，该系统用于韩国最大的互联网搜索引擎。

本专访主要介绍了韩国高校的发展状况,在韩国创立新公司的艰辛，概率统计理论在数据管理技术中应用，主存数据库中的查询优化问题，以及如何不做论文发表的奴隶等内容。

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。现在我们在伊斯坦布尔–2007 年 ICDE 的主会场。我身边的这位是黄奎勇（Kyu-Young Whang），韩国高等科学技术学院计算机科学系教授以及先进信息技术研究中心负责人。在进入 KAIST 研究院之前，黄奎勇就职于 IBM T. J. Watson 研究中心。他的研究兴趣十分广泛，包括各种各样的数据库系统，数据挖掘以及数据流。他是 VLDB 杂志的主编，前 VLDB 基金会成员以及 IEEE 会士。现在，欢迎黄奎勇。

玛丽安·温丝特：您为什么要离开企业走进高校呢？

黄奎勇：实际上，我非常享受和 IBM 那些非常优秀的伙伴一起工作的时光。但是我觉得我应该为亚太地区数据库学术界以及提升该地区数据库技术水平尽一份力。当然能和 KAIST 优秀的学生一起工作是最重要的原因。

玛丽安·温丝特：最近，我们发现有许多学者因迫切想要接触到真实的数据而离开学校进入企业。假设今天您是 IBM 一位年轻学者，您还会到学校去吗？

黄奎勇：我会的。实际上，我进入学校的动机是为了教出优秀的学生，为了能看到他们为社会做出贡献，同时也为了做自己的研究和贡献。我想让我的学生具备世界上最顶级的大学所具有的能力和竞争力。我认为有些同学是具备这样的潜力的。

玛丽安·温丝特：既然您在斯坦福生活过，那么您应该对美国高校的生活很熟悉。作为一个高校的数据库学者，您认为韩国和美国高校学术环境有什么不同？

黄奎勇：关于这个问题，我想谈两个方面。一方面是信息的获取以及获取的时效。另一方面是支持学术研究的基础建设。在 20 年前，在韩国以及亚太地区的其他国家，信息的获取是个大问题。可能在几个月之后甚至几年之后才能获得最新的信息。假设在美国开一个学术会议，可能 3 个月之后才能在韩国看到论文集，还需 6 个月才能在国内传播出去。但现在，技术的发展使得信息可以立即就被传播出去。这个问题也就迎刃而解。

然而，支持学术研究的基础设施仍然存在很多问题。基本上，韩国的学者都非常忙。主要原因是基础设施不完善，包括管理支持、技术支持以及其他方面。不断变化的评估体系以及迅速改变的社会经济体系等使得学者投入到研究的时间非常有限。我们需要为老师和学生提供一个安定的研究环境，使他们能够做出更大更长远的研究贡献。

玛丽安·温丝特：您所做的研究最主要的贡献和影响体现在哪些方面？

黄奎勇：我想提三个重要的贡献。第一个是最先在数据管理领域引入概率统计理论。数理统计就是在线性时间内，在一定误差范围内，在一个或多个属性上统计一个值的数量。能够控制误差的边界是一个非常重要的问题。这是一个令人惊讶的结果，因为一般人们都认为统计数量需要排序，这是个非常耗时的操作。现在数理统计已经应用到 IBM 的 DB2 以及其他领域，例如近似查询、数据挖掘、抽样和数据流。

在 1981 年 IBM Almaden 研究实验室（现在称为圣何塞研究实验室），我同莫顿（Morton Astrahan）和马里奥（Mario Schkolnick）一同发起这项研究。我们提出三个算法。奈杰尔·马丁（Nigel Martin）、马克·维格曼（Mark Wegman）以及菲利普·弗拉若莱（Philip Flajolet）加入了这项研究。当时，菲利普是 IBM 访问学者，他为其中一个算法做了非常复杂的分析。这项工作发表在 1987 年的信息系统以及 1990 年的 ACM TODS 两个杂志上，都具有极高的引用率。

第二个贡献体现在主存关系数据库的查询优化模型上所做的开创性工作。OBE 是第一个功能完备的主存关系数据库系统，是由 IBM T. J. Watson 研究中心摩西（Moshe Zloof）带领的团队从 1982 起历时 3 年开发完成的。我负责实现查询优化器，其中最关键的问题是需要为主存关系数据库设计一种新的代价估计模型。对于基于磁盘的关系数据库，传统的方法是通过计算磁盘 I/O 代价或其他类似的方法来估计的。但是，在主存数据库系统中，理论上是没有磁盘 I/O 的。计算 CPU 周期似乎更可行。所以，我们需要建立一个模型来计算这些周期和估计这个代价模型。

当时我提出把系统的瓶颈作为查询代价模型的基础。换句话说，我们只需对程序做概要分析，找出运行瓶颈，如消耗时间非常多的代码段。计算出瓶颈的数量并设定不同的权重（因为存在许多不同类型的瓶颈）。这样就相当于在基于磁盘的关系数据库中计算磁盘 I/O。事实上，这些瓶颈都对应着一些重要的操作，如谓词计算、元组检索等。

这在当时是一个非常新颖的想法，它给后来其他的主存数据管理系统的代价模型带来很大的影响，如 TimesTen（由 Marie-Annie Neimat 创建的公司）。该工作发表在 1990 年的 ACM TODS 上。

第三个贡献也是最重要的贡献。我们开发了 Odysseus 数据管理系统，它将数据库管理系统和信息检索紧密地结合在一起。早在 1997 年 Odysseus 就完全实现了 DB 和 IR 的结合，这给工业界带来很大的影响。Odysseus 为我们在韩国和美国都赢得了许多荣誉。该系统核心代码包括 45 万行 C 和

C++ 高质量商用代码。在 2005 年东京召开的 ICDE 会议上，我们演示了 Odysseus 系统并获得了最佳演示奖。Odysseus 成为 NHN 公司的 Naver（全称 Navigator）最主要的搜索引擎之后，对工业界已经产生了巨大的影响。NHN 是韩国最大的、比 Google 更受欢迎的互联网门户。1997 ~ 2000 年是创业阶段，之后在 Odysseus 的帮助下，NHN 公司迅速发展成为市值 60 亿美元的公司。现在，我们知道 DB/IR 集成技术已经成为一个新的热点研究领域。

玛丽安·温丝特：您刚刚提到，在韩国最受欢迎的搜索引擎是 Naver。最初，Naver 用 Odysseus 作为搜索引擎。它的点击率和用户比 Google 多两倍。既然现在已经有了韩文版的 Google，为什么在韩国 Naver 还是比 Google 更受欢迎呢？

黄奎勇：我想是在于它的商业模型。首先，Naver 的商业模型更适合韩语环境，他们已经存储大量的韩文页面。其次，它还充分运用用户创建的内容，像博客、社区以及其他内容。Naver 已经拥有大量的忠实用户，这些用户已经在博客等地方发表了大量的内容。第三，Naver 和许多重要数据提供商有合作，包括报纸出版社。他们收集报纸并在 Naver 上提供相关服务。用户宁愿在 Naver 上读报纸，而不到报纸自己的网站上。现在，这种商业模型已经融入韩国的社会文化中。

玛丽安·温丝特：您为什么不创建一家新公司来推广 Odysseus 呢？

黄奎勇：就像我之前提到的，我认为 Odysseus 通过 Naver 已经有了很大的商业影响。是否要成立一家公司是另一个问题。我一直在试图将这一技术转移到其他公司，包括像 LG 电子这样闻名于世的韩国公司。所以，我认为研究的目的是为了提出一种技术并可以实现成果转移。创办公司是第二选择。

与美国相比，在韩国成立一家新公司并不容易。有许多问题需要解决。没有健全的政策支持快速创建一家新公司。所以，你在新公司上必须投入大量的时间。既然我已经是一个非常忙的教授了，自然没有多余的时间去创建

一家新公司。

玛丽安·温丝特：我听说在韩国您周六晚上开博士讨论班，从晚上 8 点开始，直到凌晨 2 点以后。还听说您善于提一些很好的问题。我相信有很多读者想成为一个会提问题的人。可以谈谈您的小窍门吗？

黄奎勇：首先，我要说明一点：在韩国周六一直都要工作半天，直到几年前，政府推出一项政策将周六规定为假日。因为晚上比较安静，所以我们经常夜间工作到很晚。周六晚上是一个很好的选择，所以我们经常在周六晚上讨论。学生们也常常在那时提出一些很好的想法。但是，最重要的是如何将这些想法变成真正的好想法。我经常提醒我的学生注意两件事。一件是要通过和已有经典的方法做比较证实这些新想法。我们实验室已经积累大量的相关资料。这就像只有亲自摸过才知道一块金属是不是金子。

玛丽安·温丝特：您的意思是通过您的实验室的软件来做测试吗？或者您是在谈您的实验室的情况吗？

黄奎勇：确切地说，我并不是在谈软件。但是当你有想法时，你必须有一个验证模型。这个模型可以来自你自己的专业积累，或者来自我们实验室。所以，有了新想法，必须要和你知道已有的最经典的方法做比较。空想是远远不够的。

第二件事是你必须检验新想法的完备性，是否考虑到各种情况。完备性检验是一个非常有效发现漏洞的方法。不管是否说得通，人们总是忽略方法的完备性，只考虑常规情况。如果你检测了完备性，你就会很容易发现新想法的漏洞，这些漏洞就是很好的提问素材。

玛丽安·温丝特：您现在是韩国信息科学学会（韩国的 ACM）的主席。对于学会的未来发展，您有什么构想？

黄奎勇：能够当选韩国信息科学学会（简称 KISS）的主席我倍感荣幸。

为了和“kissing”区别，我们读作‘i’是长音的“kiss”。KISS是韩国历史最悠久、规模最大的计算机和信息相关的专业学会。因为学会为他们提供了许多参与各种各样研究性或工业导向的活动的机会，韩国本土的学者和技术人员都十分重视这个学会。对他们来说，参加国际性学术活动的机会受到两方面的限制。一方面，显然并不是所有的人都有出国交流的机会。另一方面，许多活动都是针对本国的问题。因为韩国与他国所处的发展水平是不同的，所以我们所关注的问题，我们认为重要的事情，我们所要达到的目标与其他国家都是不同的。

在我任职期间，我一直致力于加强几个方面：首先，要加强韩文版的杂志，这样许多本国的学者可以在这些杂志上发表他们的研究成果；其次，要创办英文版的杂志；接着，引入会士制度来评估学会成员的贡献；加强中小学和高中计算机科学的教育。目前韩国的高等院校正面临着学生缺乏早期计算机学科教育的问题。因此这是个很重要的问题；最后，要加强大学新生和高年级学生编程技巧的教育。当然，还有很多其他问题需要解决。KISS要成为学会成员表达想法和达到他们目标的场所和媒介。我有责任帮助他们达到目标。

玛丽安·温丝特：您认为一个成功的研究生或工程硕士教育应该具备哪些要素？

黄奎勇：教育包括许多方面，我只想谈一下博士和硕士的培养。我认为一般情况下，计算机科学尤其是数据库领域都是系统原型驱动的。尽管理论研究也很重要，但我们更关注研究的应用价值。所以，我坚信系统导向的研究更加重要。学生应该有很强的系统编程基础。当然，对博士而言，创新性能力的培养是最重要的。但是，如果有个坚实的系统编程能力，在选择研究课题时就会更自由，研究就会更具有实用性。

在这方面，我们的Odysseus项目为我的学生提供了很多这样的机会。我们已经培养出许多拥有很高编程技巧、具有国际竞争力的学生。现在，我有

两个学生在 IBM Almaden 实验室做博士后。一位和迈克尔·凯里（Michael Carey）一起开发 DB2 对象关系数据库。另一位同盖伊（Guy Lohman）和沃尔克（Volker Markl）一起研究高级查询优化技术。听说他们的工作都已经转化为产品，他们都已经做出了很大的贡献。所以，我的学生以及他们系统导向的研究能力在 IBM Almaden 实验室得到很高的评价，我为他们感到非常骄傲。

玛丽安·温丝特： 现在，您的儿子已经是斯坦福数据库组的一员。您对他和您从事相同行业有什么看法？

黄奎勇： 史蒂文（Steven）能成为斯坦福计算机学院的博士生，这是一件很棒的事。我非常高兴并为他感到骄傲。在那里有很多艰巨的挑战在等着他。在斯坦福大学求学是史蒂文的梦想，现在终于美梦成真了。事实上，他就是在斯坦福出生的。

做什么研究和从事什么行业这些完全取决于史蒂文自己。很多年前，我曾建议他读一个医学博士，但显然他不喜欢，他更喜欢计算机科学。既然他那么想成为一个计算机科学家，我并不介意他和我从事相同行业。这个行业会面临很多挑战、提供许多创新的机会并会对现实社会产生很大影响。我还会毫不犹豫地告诉他现在乃至以后数据库领域都是一个非常重要的研究领域。因为在信息时代，如何处理和充分利用这些海量数据是个至关重要的问题，我相信这会持续很多年。

玛丽安·温丝特： 对于那些刚入行或者资历尚浅的学者和技术人员，您有什么建议吗？

黄奎勇： 有许多建议。但是，我要强调的一点是：你一定要有一个长远的目标和方向。当然，计划可能会变，有时必须要变。但是，你必须有自己的方向和构想。否则，你很容易受不停变化的热点研究和技术的影响，迷失方向。许多人尤其是学术界的人们很容易变成出版的奴隶，他们所做的工作

就是如何写一篇论文。我认为这不是一个好方法。所以，你需要有一个研究构想，一个坚定不变的研究方向，并且你必须持之以恒。最后，你会发现许多形式不同的研究（包括一些热点研究）都会归入你的研究领域。这会更坚定你的构想。

玛丽安·温丝特：假设您有足够多的时间让您可以做一件现在无法做的事情，您想做什么？

黄奎勇：我想去做运动，锻炼身体。现在我都没有好好地锻炼身体，将来我会加强锻炼。我最喜欢的运动就是在我家附近的山上散步，今后我会经常去。

玛丽安·温丝特：作为一个计算机科学家，如果您能够改变关于自己的一件事情，您想改变什么？

黄奎勇：我想要改变很多事情。如果一定要指定一件，我想改变沉溺于工作和成就的现状。人们称我们这类人为工作狂。我想这并不仅仅是计算机科学家的问题，当今社会上有许多工作狂，他们会牺牲掉自己的生活。计算机本来是为了帮助人们减轻工作的负担，但实际上适得其反。当我们拥有更多先进的技术、更多的计算机时，我们会工作更久想要获得得更多。所以，将来我想在这方面改变自己，我将花更多的时间陪我的家人和越来越老的父母，还要常常和朋友聚一聚。

玛丽安·温丝特：非常感谢您能接受我们的采访。

黄奎勇：谢谢！

（富丽贞 译，孟小峰 审校）