数据思维应成为计算思维不可或缺的组成部分

杜小勇
中国人民大学

关键词：数据思维　第四科学范式

在8月10~17日举行的CCF计算机课程改革导教班上，中国科学院大学教授徐志伟开设了“计算机科学导论”课程，使用的是他和中科院计算所研究员孙晓明合著的同名教材[1]。课程的核心是介绍“计算思维”及教学实践。他强调，计算机科学解决问题最本质的方法就是计算思维，包括逻辑思维、算法思维、网络思维和系统思维。这四个方面也构成了该书最核心的四章。

尽管上述四个方面都是计算思维的重要组成部分，但是笔者认为《计算机科学导论》一书忽略了一项重要的内容，即数据思维。数据思维也是计算思维不可或缺的组成部分，但该书仅在第五章系统思维之下，用不到6页的篇幅讲述了关系数据库相关的两个创新故事和一道习题。本文从数据思维的内涵、独特性和重要性三个方面加以阐述。

何为数据思维？可以从认识世界（构建模型）和改造世界（解决问题）的角度来理解。从数据的角度理解世界，世界是由实体和联系构成的，这些实体和联系可以使用E-R模型(entity-relationship model)等来刻画。据此，我们可以在数字世界里用数据来表达物理世界的对象和联系。也就是说，在数字世界中，可以有一种数字孪生体存在，物理世界对象的变化可以以数字的形式在数字世界中反映出来。因此，人们可以在数字世界里用工具去探索和认识物理世界，发现其规律，或者构建机器学习模型，去预测物理世界的变化趋势等。这种认识世界（用数据构建认识世界的模型）和改造世界（通过数据探索寻求解决问题的办法）的方法就是数据思维。学术界将这一方法称为第四科学范式，这是相对其他实验观察、理论推导和计算机仿真这三种传统的科学研究范式而言的。

数据思维与逻辑思维、算法思维、网络思维和系统思维相比具有明显的不同点。首先，逻辑思维强调的是解决问题的逻辑性和正确性，算法思维强调的是计算过程，网络思维强调的是关联，系统思维强调的是整体性。而数据思维强调的是物理世界和数字世界的反映性，数据是对物理世界的反映。其次，数据思维强调“依数据说话”,基于数据本身解决问题，因此，在解决问题时会先收集数据，然后在数据上探索数据，发现解决问题的途径。这与逻辑思维强调推理明显不同。这就是数据思维的独特性。

世界进入了数字时代，利用大数据解决复杂的问题已成为一种共识。物理世界的问题非常复杂，特别是人参与的社会系统的问题更为复杂，目前还缺乏有效的数学工具将其模型化。让数据成为物理世界的一种模型，在缺乏因果关系的情况下，发现一些关联关系，能部分地解决物理世界的难题，改善人们的生活。大数据在各行各业中的应用已经看到成效，可以相信，数据思维的重要性将会越来越明显。

按照徐志伟老师的《计算机科学导论》的范例，

每一种思维方法都包括三个部分：一个实例、思维要点和创新故事。笔者也将从这三个方面讨论数据思维。

1. 从一个实例看数据思维

【例】航空联程设计：任意给出旅行的起点和终点，如何给出一个行程建议，使得在某些指标上“最短”？如果用传统办法，这是一个典型的图的最短路径问题，图的顶点就是机场，两个顶点之间的边对应两个机场之间的距离。这个问题可以用 Dijkstra 算法或者动态规划算法来求解，算法复杂度为 $O(n^3)$，其中 n 为图的顶点数。这个算法的复杂性对于大图来说还是有点高，如果节点数量过高，普通的服务器可能就算不出来了。对于这个问题，我们还可以采用数据思维来求解。我们可以记录物理世界中人们旅行的选择，构建旅客、机场以及旅客航程关系的数据模型，这是旅行大数据。我们只要根据此旅行大数据先搜索出全部的从起点到终点的旅行历史记录，然后用简单的统计方法，对最受欢迎的路线进行一个排序。这个结果就可以看作是过去旅客的经验选择，完全有理由推荐给客户，供他们选择使用。这样做的算法很简单。当然，我们还可以在旅行大数据上利用更复杂的数据挖掘和机器学习方法，来发现和探索规律，改进服务。

2. 数据思维的要点

利用数据思维来求解问题的过程，包括以下几个要点：第一，数据采集与汇聚。为了解决复杂问题，首先需要采集能记录复杂问题所涉及的物理世界中有关对象实体和联系的数据，然后对采集的数据进行整理。第二，为了方便数据的使用和探索，需要对数据进行建模、组织和管理，各种数据管理工具应运而生。第三，数据分析与数据挖掘。探索数据是数据思维的基本活动，包括开发数据分析和数据挖掘软件发现数据中隐藏的规律，或者利用数据训练模型。尽管开发软件需要逻辑思维和算法思维等，但是数据思维体现的是一种黑客似的探索思维。第四，数据可视化。向用户提供图形化的结果展示工具。这种直观的展示分析结果，有利于人们理解结果、发现规律。

3. 数据思维的创新故事

利用大数据进行创新的故事很多，各行各业都有。例如，吉姆·格雷 (Jim Gray) 通过将天文望远镜所拍摄的图片组织起来放到互联网上供网民访问，改变了天文学的研究范式 [2]；利用 Freebase 等开放资源获得超过 15 万欧洲历史名人的数据，可以从这些人的出生地和死亡地分析发现欧洲历史上“条条道路通罗马”的事实 [3]。还有很多其他例子，例如通过基因测试收集人类基因数据，拓展了生物科学研究新领域；通过监控用户上网行为，对用户进行画像，并预测用户的行为，开发了精准营销新模式等。

因此，在大学计算机教育中，加强数据思维的培养非常重要。根据我们对数据思维的理解，大学计算机、数据科学概论、数据库系统概论，乃至数据分析和数据挖掘等课程都是培养学生数据思维的最好载体。笔者也在今年的 CCF 计算机课程改革导教班上讲授了一门课“数据库与大数据”。这门课的教学目的就是要让学生感受并掌握数据库方法。数据库方法内容丰富，数据思维就是其中最为重要的内容之一。这门课也是实践数据思维很好的途径。 ■

致谢：本文在成文过程中得到北京大学教授李晓明、中国人民大学教授王珊和《计算机科学导论》的作者徐志伟教授等人的积极鼓励和修改建议，一并表示感谢。

杜小勇
CCF 常务理事、会士、CCF 教育工委主任。中国人民大学教授。主要研究方向为数据管理、智能信息检索。
duyong@ruc.edu.cn

参考文献

[1] 徐志伟，孙晓明 . 计算机科学导论 [M]. 清华大学出版社，2018.

[2] Hey T, et al. *The Forth Paradigm: Data-Intensive Scientific Discovery*[M]. Microsoft Research, 2009.（中译本：潘教峰等译 . 第四范式：数据密集型科学发现 [M]. 科学出版社，2012）

[3] Schich M, et al. A network framework of cultural history[J]. *Science*, 2014, 345(6196): 558-562.