10个关键计算机项目改变科学进程

■任　天

从Fortran到arXiv.org,这些计算机编码和系统让数学、生物学和物理学等学科的发展达到了真正“日新月异”的速度。如今,科学研究从根本上已经与计算机紧紧联系在一起,并渗透进了研究工作的各个方面。近期,《自然》(Nature)杂志也将目光投向了幕后,针对过去60年来改变科学研究的计算机代码,盘点出了10个最为关键的项目。

语言先驱:Fortran编译器(1957年)

最初的现代计算机并不容易操作。当时的编程实际上是手工将电线连接成一排排电路来实现的。后来出现了机器语言和汇编语言,允许用户用代码为计算机编程,但这两种语言都需要用户对计算机的架构有深入的了解,这使得许多科学家难以掌握。

直到20世纪50年代,随着符号语言的发展,特别是“公式翻译”语言Fortran出现后,这种情况发生了改变。利用Fortran,用户可以用人类可读的指令来编程,然后由编译器将这些指令转换成快速、高效的机器代码。Fortran至今仍然广泛应用于气候建模、流体动力学、计算化学等学科,这些学科都涉及复杂线性代数并需要强大的计算机来快速处理数字,而Fortran生成代码的速度很快。

信号处理器:快速傅里叶变换(1965年)

当射电天文学家扫描天空时,他们捕捉到的是随时间变化的复杂信号杂音。为了理解这些无线电波的本质,他们需要看到这些信号作为频率函数时的样子。一种名为“傅里叶变换”的数学过程可以帮到研究人员,但它的效率很低,计算过程也十分复杂。

1965年,美国数学家詹姆斯·库利和约翰·杜基想出了一种加速该过程的方法,即快速傅里叶变换。快速傅里叶变换通过递归(一种通过重复将问题分解为同类的子问题而解决问题的编程方法)将过程进行了简化。于是开启了快速傅里叶变换在数字信号处理、图像分析、结构生物学等领域的应用,成为应用数学和工程领域的重大事件之一。

分子编目:生物数据库(1965年)

如今,科学家所用的庞大基因组和蛋白质数据库都源于美国物理化学家玛格丽特·戴霍夫的工作,同时,她也是生物信息学领域的先驱。20世纪60年代初,当生物学家们致力于梳理蛋白质的氨基酸序列时,戴霍夫开始整理这些信息,以寻找不同物种之间进化关系的线索。她与3位合著者于1965年发表了《蛋白质序列和结构图谱》,描述了当时已知的65种蛋白质的序列、结构和相似性。历史学家布鲁诺·斯特拉瑟在2010年写道,这是第一个“与特定研究问题无关”的数据集,它将数据编码在打孔卡中,这使得扩展数据库和搜索成为可能。

预测领先者:大气环流模式(1969年)

在第二次世界大战结束时,计算机先驱约翰·冯·诺伊曼开始将用于计算弹道轨迹和武器设计的计算机转向天气预测问题,试图基于物理定律进行数值天气预测。在早期预测天气时,气象学家需要输入当前的条件,计算它们在短时间内会如何变化,并不断重复。这个过程非常耗时,以至于在天气状况实际出现之前还无法完成数学运算。而计算机就能让这个问题变得容易。

20世纪40年代末,冯·诺伊曼在普林斯顿高等研究院建立了天气预报团队。1955年,他的第二个团队——地球物理流体动力学实验室,开始进行“无限预测”,也就是气候建模。1969年,他们创造的海洋—大气联合模型划分的面积为500平方公里,将大气分为9个层次,虽然只覆盖了地球的六分之一,但这使研究团队第一次能够通过计算机预测二氧化碳含量上升对气温的影响,成功创造了科学计算“里程碑”。

数字运算机:BLAS(1979年)

科学计算通常涉及使用向量和矩阵进行相对简单的数学运算,但这样的向量和矩阵实在太多。在20世纪70年代,甚至没有一套普遍认可的计算工具来执行这些运算。因此,从事科学工作的程序员会将时间花在设计高效的代码来进行基本的数学运算,而不是专注于科学问题。

1979年,BLAS(基本线性代数程序集)出现了。这是一个应用程序接口(API)标准,用以规范发布基础线性代数操作的数值库,如矢量或矩阵乘法。该标准一直发展到1990年,为向量数学和后来矩阵数学定义了数十个基本例程。40多年来,BLAS代表了科学计算堆栈的核心,也就是使科学软件运转的代码。

显微镜必备:NIH Image(1987年)

20世纪80年代初,程序员韦恩·拉斯班德在马里兰州贝塞斯达的美国国立卫生研究院的脑成像实验室工作。该实验室拥有一台扫描仪,可以对X光片进行数字化处理,但无法在电脑上显示或分析。拉斯班德因此而建立了NIH Image。

NIH Image及其后续版本使研究人员能在任何计算机上查看和量化几乎任何图像,其系列软件ImageJ向外界提供了一个看似简单、极简主义的用户界面,但内置的宏记录器(允许用户通过记录鼠标点击和菜单选择的序列来保存工作流)、广泛的文件格式兼容性和灵活的插件架构,让该工具具有无限的可扩展性。这个程序的目的不是做到一切或终结一切,而是服务于用户的目标。不像Photoshop和其他程序,ImageJ可以成为你想要的任何东西。

序列搜索器:BLAST(1990年)

寻找序列之间的相似性,可以让研究人员发现进化关系,并深入了解基因功能。但在迅速膨胀的分子信息数据库中,想要快速而准确地做到这一点并不容易。

玛格丽特·戴霍夫在1978年提供了关键的进展。她设计了一种“点接受突变”矩阵,使研究人员不仅可以根据两种蛋白质序列的相似程度,还可以根据进化距离来评估它们的亲缘关系。直到1990年,研究人员们开发了一种更强大的改进技术,即BLAST。BLAST能将处理快速增长的数据库所需的搜索速度,与提取进化上更为遥远的匹配结果的能力结合起来。与此同时,该工具还可以计算出这些匹配发生的概率,并且很容易使用。

预印本平台:arXiv.org(1991年)

20世纪80年代末,高能物理学家经常将他们已投稿的论文手稿副本邮寄给同行,以征求他们的意见。1991年,当时在新墨西哥州洛斯阿拉莫斯国家实验室工作的金斯帕格编写了一个电子邮件自动应答程序,希望建立一个公平的竞争环境。订阅者每天都会收到预印本列表,只需通过一封电子邮件,世界各地的用户就可以从实验室的计算机系统中提交或检索论文,并获得新论文的列表,或按作者或标题进行搜索。

金斯帕格原本的计划是将论文保留3个月,并将内容限制在高能物理学界。但一位同事说服他无限期地保留这些文章。他说:“就在那一刻,它从布告栏变成了档案馆。”于是,论文开始从各个领域如潮水般涌来。1993年,金斯伯格将这个系统迁移到互联网上,并在1998年将其命名为arXiv.org,沿用至今。

数据浏览器:IPython Notebook(2011年)

Jupyter notebook对于初学Python的人来说简直是一个无比友好的存在,它能显示Markdown,也能显示并执行代码,比传统的交互式和全脚本执行更加清晰和简便,而它的前身就是来自费尔南多·佩雷斯发明的IPython Notebook。

2001年,正在读博的费尔南多·佩雷斯认为,Python并不是为科学目的而构建的。例如,它不允许用户方便地预加载代码模块,也不允许打开数据可视化。因此,佩雷斯自己编写了另一个版本,创造了IPython。这是一个“交互式”Python解释器,由佩雷斯在2001年12月推出,共有259行代码。10年后,该工具迁移到web浏览器上,推出了IPython Notebook,开启了一场数据科学革命。

快速学习器:AlexNet(2012年)

人工智能有两种类型:一种是使用编码规则,另一种则通过模拟大脑的神经结构来让计算机“学习”。在ImageNet比赛中,研究人员被要求在一个包含100万张日常物体图像的数据库中训练人工智能,然后在一个单独图像集上测试生成的算法。当时最好的算法错误分类了大约四分之一的图像。而AlexNet是一种基于神经网络的“深度学习”算法,它将错误率降低到了16%。

加拿大多伦多大学的计算机科学家杰弗里·辛顿表示,AlexNet的成功反映了足够大的训练数据集与出色的编程,以及新出现的图形处理单元的强大能力的结合。“突然之间,我们可以将(算法)运行速度提高30倍,”他说,“换句话说,我们可以学习多达30倍的数据。”