量化不可避免，避免简单量化

张 健
中国科学院软件研究所

关键词：科学成果评价 量化

近年来，关于如何对学术成果以及对科研人员进行评价，已经有了很多讨论。国家有关部门也在逐步制定比较合理的政策。总的趋势是，重“质”不重“量”。

2015年7月，在中国计算机学会(CCF)学术工委内部讨论期间，笔者曾在邮件里写过：“**量化不可避免；但避免简单量化**”。这里，笔者稍微展开一下，谈谈自己的一些想法。

量化不可避免

在学术界，我们免不了各种评审，比如评职称、评项目、评奖等。评审的结果基本上可以抽象为数字。比如若干教授去评级，结果可能是张三升为三级教授而李四还是四级教授;又如每年有很多人（团队）去报奖，有的得到一等奖、二等奖，有的则名落孙山。有些单位把发表几篇（一定档次的）论文作为研究生毕业、教职工提职称的先决条件。

从这个角度看，某种形式上的量化，不可避免。我们总是需要把模糊的科研成果与贡献映射到数字（有点类似于模数转换）。可是，大家又经常听到、看到这样的观点：不要去“数”文章，要重“质”不重“量”等等。似乎“量化”是一件不好的事。

笔者的理解是，“量化”有多种形式，不宜一刀切完全抛弃，只是简单、粗暴的“量化”不太好。

简单量化

所谓的“简单量化”，在不同时代，可能会有不同的具体做法。

最简单的做法是数文章数量。当然，还可以进一步细化。比如，十多年前，可能主要是统计SCI文章数量。最近几年，在计算机界越来越多的人统计CCF-A类[1]文章的数量。更进一步，还可以对不同层次的文章加权处理。比如，A类文章或SCI一区文章，5分；B类文章3分，等等。

简单量化的弊病，大家都知道。一个好的成果不一定发表在顶级期刊、会议上；发表在顶级期刊、会议上的论文，并不一定都是好的成果。

除了统计文章数量以外，成果被引用次数以及相关的量化指标（如h-index）也逐渐被很多人使用。相对而言，这种方式稍微复杂一些，成本要高一些。至少，要有人统计文章间的引用关系。

如果不用简单量化的方法，随之而来的一个问题是，应该采用什么方法呢?

科研成果形式的多样性

2002年初，笔者曾写过一篇短文，谈了笔者关于计算机领域评价体系的一些想法(见侧栏,有删节)。

笔者觉得，科研成果的形式有很多种，包括：论文、著作、(软件)原型系统、语言、标准、专利，

[1] CCF推荐国际学术会议和期刊目录，分为A, B, C三类。

等等。论文之外的具体成果，如 TeX，Linpack 以及各种程序设计语言。

既然科研成果的形式各种各样，那么其量化手段也应该更加丰富。除了前面讲的论文发表处、论文之间的引用关系以外，还可以包括：

- 被解决的理论问题在此之前开放 (open) 了多少年；
- 所提出的程序设计语言或其他描述语言有多少人使用；
- 所提出算法的复杂度；
- 软、硬件系统的性能（相对于一定的 benchmarks）；
- 软件的（重要）用户数、被下载次数，是否开源、开源社区的活跃程度；
- 专利转让带来的直接经济效益，等等。

从相关学科看，还有其他的量化方式。比如，在国际数学家大会上做一小时或 45 分钟报告[2]。

如果只针对一种形式的成果（比如发表论文）制定简单的鼓励数量的政策，可能会有副作用。在这种政策下，人们可能会想，与其花很多精力去做一个（基本）可用的原型系统，还不如去多写几篇文章。这样的导向会减缓技术转移的步伐，使得学术界和产业界的鸿沟难以逾越。

内涵式的量化

除了表面的数字以外，笔者个人更赞同对科研成果的实质性内容进行量化。

计算机科学作为一门技术学科，其很多成果是可以改变世界的（而自然科学以认识世界为主）。对很多技术成果的价值，可以采用 weight × delta 这样的公式来衡量。这里的 weight 是指被研究问题的重要性 (significance, relevance)，delta 是科研人员所做的对现状的改变。

举例来说，对同一个难解的问题 (intractable problem)，如果 A 作者给出的方法可有效解决 70% 的问题实例，B 作者给出的方法可有效解决 20% 的问题实例，那么，粗略地看，可认为 A 的贡献更大。当然，如果细分，还可以考虑不同实例具有不同的难度、不同的价值。

如果 C 作者发表了一篇货真价实的论文，大大地推进了当前的技术水平 (state of the art)，而 D 作者发表了若干篇微小改进的论文（least publishable unit 或 minimum publishable unit），那我们应当给出恰当的分值，以体现前者比后者贡献更大。

2014 年，在一个关于“十三五”规划的小范围讨论会上，大家回顾了过去几年的学科进展。在算法和复杂性方面，有学者提到该领域近期的一个显著进展是关于矩阵乘法复杂度的认识：国外几位学者先后将其降为 $O(n^{2.374})$，$O(n^{2.373})$，$O(n^{2.3729})$。在此之前，最好的结果是 $O(n^{2.376})$(Coppersmith-Winograd, 1991)。虽然指数的改变十分微小，但考虑到该问题的基础性、重要性及难度，仍然可以把上述成果看成是重要进展。这可以当作 weight 很大、delta 较小的一个例子。当然，要采纳上述手段来量化并不太容易。对某些前瞻性研究成果，暂时可能还难以准确估计其价值。要有小同行作为评委，要能把握当前和过去的技术水平，这样才能给出 delta。也许，有人觉得，这种方式不够“客观”。但是，我们在采用简单量化的方法时，它的一些基础数据（发表了几篇论文），其实也依赖于有关会议或刊物的审稿人的“主观”意见。

对于如何量化 (weight) 研究问题的重要性，不同的人也会有不同的见解。希望学术界能在不同层次总结一些公认的重要问题列表（类似于数学界的希尔伯特问题[3]或者 Millennium Prize Problems[4]），产业界能提出一些“卡脖子”问题（这里说的“卡脖子”并不一定针对某一个国家）。事实上，几年前，教育部等单位组织了《10000 个科学难题》的征集，

[2] 国际数学家大会 (International Congress of Mathematicians, ICM)，是由国际数学联盟 (IMU) 主办的国际数学界规模最大也是最重要的会议，每四年举行一次。它是全球性数学科学学术会议，被誉为数学界的奥林匹克盛会。每次大会一般都邀请一批杰出数学家在大会上作一小时的学术报告和在学科组的分组会上作 45 分钟学术报告。

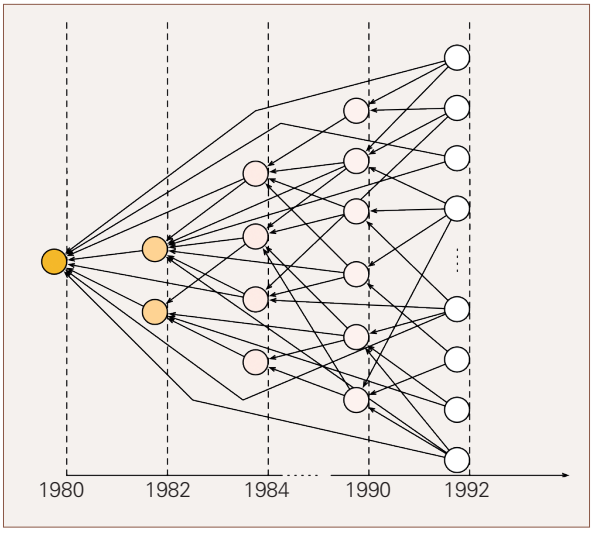

图1 领域发展历程图

笔者作为信息科学卷的编委，也参与了此活动。今年，有关单位开始梳理产业界的一些瓶颈。

源头创新

我们不时听到，要重视“源头创新”。但是，简单量化的做法有可能与此目标相悖。

一个领域（研究方向）的发展历程如图1所示。在开始阶段，发表的论文可能很少，但很重要。在繁荣期，论文数量很多，被引用次数可能也很多。但其平均价值未必很高。

评价一个成果，是属于无中生有、创立新学说，还是发扬光大、助力学科成长，抑或只是添枝加叶、略增学科体量？是改变主流研究方向，还是随波逐流？要给出比较准确的回答，还是需要内行人士厘清学科发展的脉络，特别是flow of ideas，进而看出谁做了什么重要贡献。对源头创新成果的评价，可能需要比较长的时间，在短期考核的机制下，也许难以显示出其价值。

多作者贡献的分配

一篇论文有多个作者，是很常见的。在有的领域（如，理论计算机科学），很多论文的作者是按照姓名的字母顺序来排。在计算机界的其他领域，则可能是按贡献来排序。但是，每个作者的具体贡献程度，也很难量化。

国内目前的现状似乎是，只有第一作者和通信作者“算数”。这也不尽合理。一方面，其他作者可能在某个局部工作上起了关键作用；另一方面，有的通信作者，其实质性贡献也可能很小（很多国际会议论文并不标注通信作者）。

小结

在计算机界，科研成果的形式多种多样，对成果的量化评价，也应该有多种方式。

与简单量化的做法不同，笔者更希望大家从解决问题的角度、从引领学科发展的角度来进行评价。希望国内同仁能取得更多经得住时间考验、经得住实际检验的科研成果。 ■

（写于2018年9月26日）

致谢：

如果没有钱德沛、包云岗等编委的热情鼓励，本文不可能完成。本文写作过程中，得到了王戟、许可、孙晓明、晏荣杰的帮助。在此一并致谢。

张 健

CCF杰出会员，CCF学术工委执行委员、公共政策委员会执行委员、专委工委委员。中国科学院软件研究所研究员。主要研究方向为自动推理、约束求解、程序静态分析与检错、软件测试数据生成。zj@ios.ac.cn

[3] 希尔伯特在1900年8月巴黎国际数学家代表大会上，发表了题为“数学问题”的著名讲演。他根据过去特别是19世纪数学研究的成果和发展趋势，提出了23个最重要的数学问题。

[4] 千禧年大奖难题，又称世界七大数学难题，是七个由美国克雷数学研究所(Clay Mathematics Institute, CMI)于2000年5月24日公布的数学猜想。

侧栏

关于计算机领域基础研究评价体系的一些想法

张 健
中国科学院软件研究所

2002年1月

评价体系应与具体的学科领域相关。根据我自己的专业，我只能谈一些计算机方面的情况。我觉得计算机学科大致还是属于技术科学。技术科学的主要任务是改变世界，促进生产力发展。当然，也有一部分计算机科学研究可看成是探索规律性质的自然科学研究，而且知识创新和技术创新也不是截然分开的（比如相对论和原子弹之间密切相关）。尽管如此，我仍然从技术进步的角度来讨论。

通常，评价分为定性和定量两类。但是量化指标不一定就是指论文篇数以及所发表刊物的影响因子等等。还可以包括其他一些，如性能指标。例如，解决某问题原来所用的时间为1小时，后来某人提出改进措施，使解题时间变为1秒。那这就是相当大的进步。吴文俊先生提出的几何定理证明方法影响很大，主要原因就是原来的一些方法不太有效，而吴方法能很容易地证明大量定理。没有多少人关心他的文章是在哪里发表的。

当然从可操作性来看，上述指标不一定总是可行。有些基础研究成果离实用还远。它究竟是否有用要等将来才知道，现在还不好讲。这时就只能看专家或同行意见（包括文献中对该成果的评价）。当然，这难免会有不同看法（仁者见仁，智者见智）。而且有时找不到小同行，这时只能看其论文发表情况。不过论文并不一定越多越好。如果只有一篇论文，但它确实能很好地解决问题，这也比若干篇不疼不痒的论文好。YACC（编译器的编译器）就是一个好例子。我欣赏这种一步到位、一针见血的工作。

总之，我觉得在评价研究成果时可以依次考虑下述因素：对当前技术水平 (state-of-the-art) 的推进程度以及这种技术的重要性；被同行引用和评价情况，被收录情况；论文所发表的刊物/会议录的水平以及论文篇数。

除了论文以外，原型系统也可看作成果的一种形式。它也不一定就要申请专利。它的重要性在于，它表明哪条路走得通，哪条路走不通（或者有什么困难）。做原型系统可以避免“不着边际”的研究，也可以扩大影响。但在国内很多情况下，原型系统往往只是给论文提供一些支持，或在鉴定时给专家看一看。

对国内研究的一个不太好的印象是，“顶天”与“立地”分开，自然科学奖与科技进步奖离得很远（前者强调深刻的理论结果而后者强调经济效益）。很多人把研究等同于理论研究，我不赞同。如果我们国家有人发明了Unix或Linux这样的系统，该给他什么奖？我的上述印象不一定准确，因为我自己很少参加评审、鉴定这类活动，对我国研究的整体水平和状况缺乏足够的了解。

当然，在建立评价体系时必须考虑到我国国情（国内当前的科研水平）。否则可能不切实际。另外，还要考虑可操作性。跨学科之间的评价比较难。