让社会科学插上信息技术的翅膀

李晓明，北京大学信息科学技术学院
祝建华，香港城市大学英文与传播系

随着信息技术的普及，以人类社会活动中所产生的各种信息的数字化、网络化为基础，以越来越庞大的上网人群为对象，社会科学的发展正面临着一种诱人的新机遇。对这种机遇的及时把握，不仅有望给社会科学带来突飞猛进的发展，也预示着计算机技术（信息技术）的应用与进步有了一个激动人心的新方向。

从支持科学研究的工具的角度来看，众所周知，计算机技术给自然科学与工程科学研究带来了第三种途径和方法，即在理论研究和实验研究之外的模拟研究方法。特别是高性能计算和科学数据可视化技术，其效能和不断降低的成本使得计算模拟在各种研究活动中的角色越来越重要，有些场合甚至已是不可替代（例如核爆炸模拟）。几乎每一个学科，前面加上“计算”二字，就形成了一个专门的分支（例如计算物理、计算化学、计算力学、计算生物学，等等），在学科发展中发挥着独特的作用。

对于社会科学研究而言，计算机技术作为一个支撑工具，传统上不像在自然科学和工程科学中那么突出和广泛。但十年前兴起的World Wide Web*，以及随之而来的社会信息数字化、网络化的浪潮，不仅给社会科学带来了一个新的研究对象，而且其中气象万千、随历史长河奔流的海量数据（在很大程度上是我们社会现实的反映），以及各种新型的工具（例如搜索引擎和数据挖掘），也给研究社会科学的传统问题带来了新的机会和方法。然而，我们也看到，目前社会科学研究人员利用信息技术开展工作的现状与信息技术以及社会数字化信息不断丰富积累带来的潜力之间，还存在一个很大的空间。这个空间既表现在大量的数据有待人们应用新的社会科学方法来挖掘和开发，也表现在信息技术为适应社会科学的需要还应积极参与并发展出相应的新工具。

社会科学，这里特指“实证社会科学”（empirical social science），大致包括经济学、管理学、社会学、新闻传播学等领域，强调通过调查或者统计等手段进行数据的采集和分析，从中得出结论或者答案。就目前应用信息技术的情况看，在数据分析阶段自从有了 SPSS 等软件后，研究效率产生了一个飞跃，但在数据采集阶段，还基本上停留在手工作坊的水平。大量的社会调查（无论是面访还是通过电话），不仅耗时，而且还有很高的成本，一般要占用整个研究项目中三分之二到四分之三的资源。这对匮乏研究经费的年轻教师和学生设置了难以逾越的门槛，成为影响中国社会科学实证研究发展的一大瓶颈。

那么，为什么现在这个瓶颈有可能被舒缓，从而使得社会科学研究能得到继 SPSS 之后的另一个飞跃呢？

---

* 后面也称“互联网”。在本文中，我们对“World Wide Web”和“互联网”不加区分。

我们看到，社会科学研究在一定意义上和自然科学一样，都是信息（数据）的加工和处理。一个重要的不同是，自然科学中用于计算模拟的原始数据经常可以通过模型来产生，相对比较单纯和容易得到，例如要模拟一个飞行器的流场，可以通过飞行器的外形模型来确定周围的空间数据；相比之下，无论是搞经济模拟还是政策模拟，实证社会科学研究的原始数据都要花很高的成本才能得到。但现在，我们看到了已经开始，并还在一浪高过一浪的社会信息广泛数字化、网络化的洪流，看到了通过各种网络互联起来的人群已经增长到了一个足够充分的数量。从程控交换机开始，到互联网，到手机，到数字电视，到形形色色的传感器网络，人类的活动，以从未有过的一种广度和深度被及时地记录、传输和存储。在互联网上，从最开始“一对多”单向信息发布的网页，到“多对多”的 BBS，论坛，博客，维基。新产生的信息差不多都是与生俱来数字化、网络化的，以前的许多有意义的资料信息（例如图书、报纸）也在被重新数字化、迟早会网络化。这对于社会科学来说，预示着极大幅度提高数据采集效率的潜力。而巨大的上网人群也给社会科学家进行从未有过广泛且高效的调查提供了天然的基础。采用信息技术，我们不仅能主动进行海量信息获取（例如大规模搜集网页），还可以对隐形产生的海量信息进行知识挖掘（例如对大型网站的访问日志进行挖掘，研究网民的行为模式）。

当然，海量的数字化、网络化信息带给我们的潜力并不意味着实际利用没有困难。可能性与现实性之间依然存在相当距离。当前主要存在着抽样问题，噪音问题、分享问题、法规问题、伦理问题、终端工具（数据变换和显示）问题等。抽样问题涉及如何确保计算机从互联网上自动搜取的数据对特定研究总体的代表性。由于自发、分散、匿名等特点，目前常见的通过自动搜集获得的网站、网页、网民数据，都是“方便样本”(convenient samples)，而不是“随机样本”（random samples）。其次，即使数据的代表性良好，这种数据一般含有大量的噪音，如因转载而重复出现的内容（中文网页中冗余率为 400％之多）、随意创建而出现的无意义、人为或机器造成的错字或错误内容等，均需要先加以过滤后才能使用。此外，网上很多（甚至是大部分）数据还不是公开的，如企业、学校、机关内部论坛内容；网站用户使用行为记录；等等。其拥有者基于商业机密等考虑，一般不愿与社会科学家分享这些极有学术价值的数据。再者，即使有些数据拥有者免费或收费开放这些数据，因其中涉及隐私问题还会受到法律或社会伦理的限制。最后，即使有了大量可靠和有用的数据之后，目前还匮乏通用的终端工具，以便社会科学家们在没有任何技术人员支持下能够直接使用这些数据。

上述种种障碍，并非无法解决。抽样或噪音问题，同时涉及技术和人的因素，需要信息技术和社会科学携手攻关。如互联网信息抽样问题，与生产过程或质量控制的抽样不同，反而与社会调查的抽样更相似。噪音问题则直接倚赖于对文本语义及其社会背景知识的理解。有些信息技术专家和社会科学专家已经发现他们之间在这些问题上有很多共同语言（当然双方往往由于术语不同而没有认识到这种合作基础）。至于数据分享问题，则需要数据拥有者与社会科学家之间通过对话和探索而形成互惠互利的机制。其实这并不是一个新问题。在数字化、网络化出现前的半个世纪，商界和学界就已经摸索出许多行之有效的游戏规则。如在美国就有数个大型的社会数据档案中心，其数据由商业研究公司提供

（当然事先将数据中的个人或公司标识过滤掉）。根据我们的经验，如果有合适的工具（如将个体数据抽象到群组、团体、地区等整体），很多数据拥有者、政府部门以及个人用户都会自然打消对商业机密或个人隐私的顾虑，而产生的数据往往已经能够满足大多数社会科学家的研究需要。而信息技术界则既有责任也有能力来开发这种工具。当然，必要的法律监管或伦理约束也是急需解决的。如在香港和其它一些地区已经制订了保障个人隐私资料的具体法规。

对于这些问题，信息技术一方面大有用武之地，另一方面也面对挑战。以互联网文本信息为例，每天搜集上千万网页已不困难，但搜集一批（例如几十万）针对某个研究目标具有科学代表性的网页则不是件易事。对公开表层网页信息的搜集技术已比较成熟，但对深层网络(“deep web”)信息（即便不是被保护的）的有效搜集依然是个难题。在信息的处理和利用方面，搜索引擎的成功不仅将传统的信息检索技术带上了一个新的台阶，人工智能、并行与分布计算也在其中扮演着重要角色。那样一些曾经是阳春白雪的技术，在搜索引擎这种直接面向大众、真正大规模的应用中焕发了青春。同时我们也注意到，相比几乎无所不包的海量网络信息能够给我们带来知识的潜力，现在能用的工具还太少。例如我们想关心某个知名人士在 2005 年到达过哪些城市，某个重要事件从发生到消失在人们注意力之外的过程，在 2005 年里哪些人物或者事件最受关注，等等，都不是通过现有的工具能够容易解决的。从原理上讲，搜索引擎能够给我们的是“某一篇网页里包含的信息”，但人们很可能需要得到在一个网页集合中蕴含的信息。一些已经发展多年的技术，例如信息过滤，信息提取，文本挖掘，文本综述等，在某些特定领域的场合（例如电子商务）已经展现了成功的应用，但面对纷杂的、海量的网络信息，要回答上述类似的问题，它们在效果和效率上都还有很大的改进空间。例如现在还没有很有效的方法来获得一个网站的行政属地，而这个信息对社会科学研究是很有意义的。另外，海量信息的表现（呈现）也会是一个引人入胜的领域，信息的类型、强度、相互之间的关系等时空特性，往往只有通过适当的表现才有可能带来有效的理解 *。没有数据的可视化，各种科学与工程计算会逊色很多；没有海量网络信息的可视化，社会科学与信息技术结合也会缺少精彩。

希望信息技术能更加深入地应用于社会科学研究，也是国际国内许多学者的共识，这样的共识促进了有关机构设立了专门的计划或者项目。在英国，e-science计划之下有专门的e-social science分支，在美国的cyber infrastructure计划中也专门有对社会科学研究的安排。在中国，自然科学基金委已开始支持相关的项目 *，ChinaGrid II正在构思一个“实证社会科学网格”。一些有远见的IT企业也开始了相关技术的研发，大到IBM雄心勃勃的Web Fountain，小到现在市场上已经可以看到的诸如“企业竞争情报系统”之类

---

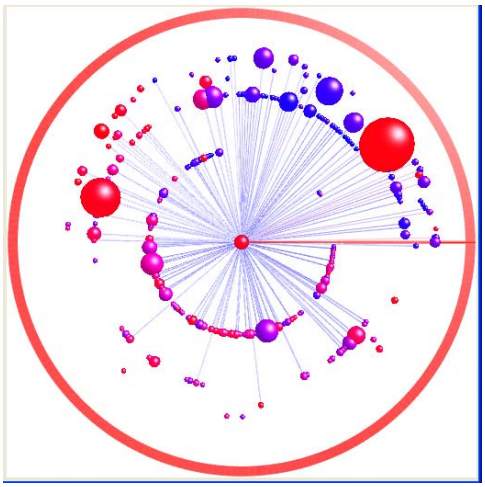
* 作为例子，附图是香港科技大学屈华民博士的研究小组和北大网络实验室合作的成果，表现的是北京大学和教育网内其他大学之间的网站互连关系等信息。图中不同元素的大小、颜色、位置、相互之间的距离都有着特定的含义。

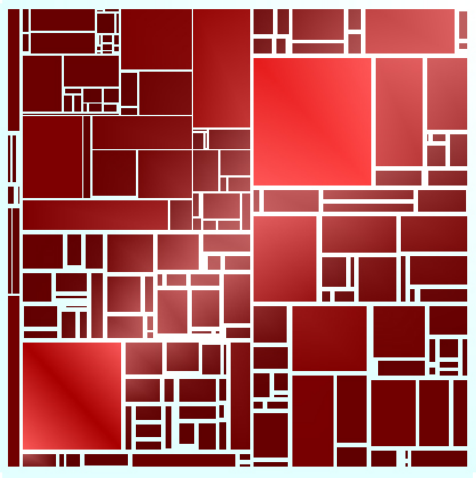
* 国家自然科学基金（60573166），Web 结构和社会信息相关性模型及其计算实验方法研究

的产品，虽然都还远未成熟，但其潜在的巨大影响已初见端倪。我们应该可以看到，与社会科学结合将是信息技术未来发展的一个重要方向（如果不是最重要的话）。

然而，在看到这种令人兴奋的前景的同时，我们也要有充分的思想准备。信息技术与社会科学的结合在技术和条件准备上还没有达到水到渠成的境地。面对动辄成百上千万的数据量（以往经常只是成百上千），新的社会科学研究方法有待诞生（例如对社会现象定义出新的评估指标），面对噪音充斥的海量信息和社会性的应用目标，新的信息处理技术也需要开发出来。特别地，从研究效益的角度考虑，有必要建立若干适应面较宽的公用社会信息基础平台或基础设施，使得各种课题的研究都可以在一个适当的基础上开展，从而降低工作的成本，激发追求成果的信心。另外，如果说计算机技术和自然科学的结合是一种天然的嫁接的话（其实也有各种各样的学科界限和心理障碍），术语的差别，思维方式的不同，将使得信息技术与社会科学的结合需要更大的耐心和努力。

现代计算机的发明已经有了六十年历史。在过去六十年里，我们不仅见证了计算机从一种尖端、昂贵、超净室里的设备变成了一种家用电器的过程，还见证了互联网将千千万万计算机互联起来，从而使得计算机本身不再是人们谈论的中心的现实，也正在见证整个社会生活经历着一个无比广泛的数字化、网络化、智能化（即自动化）的进程。这样一种摧枯拉朽、不可逆转，由信息技术带动的历史潮流对我们的影响不仅在于带来了生活的便捷，同时也为社会科学家们更细微、全面、实时地理解社会运动本身带来了“四两拨千斤”的支点和杠杆。我们预见在未来的年月里，信息技术和社会科学的相互促进会是一种日益重要的发展方向。插上信息技术的翅膀，社会科学研究将展现出一种新的面貌；应用于社会科学研究，信息技术将迎来一个新的、生机勃勃的春天。





图例：左边，是以一个机构为中心，展现它和其他机构网站的互联关系和信息强度。圆的大小表示信息量的多少，颜色表示被别的机构网站链接的程度（即“hot”的程度），和中心圆的距离表示和它链接的强度，等等；右边表示的是一个机构中各个子网站的大小和“热度”分布。

作者简介：

李晓明，1982 年本科毕业于哈尔滨工业大学计算机系，1986 年毕业于美国史蒂文斯理工学院，获博士学位。现任北京大学教授，互联网信息研究所所长。兼任中国计算机学会副理事长，中国电子学会常务理事，美国电气工程学社 Eta Kappa Nu 社员，IEEE 高级会员。目前感兴趣的主要研究领域包括海量网络信息与社会发展的互动关系，以及人工智能与并行计算技术在网络信息分析与处理中的应用。主持发展了“天网”搜索引擎(http://e.pku.edu.cn)；提出了收集存储并系统研究网页历史信息的理念，主持建设了“中国网络信息博物馆”(http://www.infomall.cn)。发表论文 100 多篇，获得 2 项发明专利，著有《搜索引擎--原理、技术与系统》(科学出版社)。目前是《中国科学》,《计算机学报》，以及《Concurrency: Practice & Experience》和《Journal of Web Engineering》等学术刊物的编委。

祝建华，美籍华裔，1982 年毕业于复旦大学获文学学士学位；1990 年毕业于美国印第安纳大学获大众传播学博士学位。曾任美国康纳迪克大学传播科学系助理教授、副教授，现任香港城市大学英文与传播学系教授，为该校传播与新媒体硕士专业创办人。目前的研究领域包括互联网、移动通讯、数码电视等新媒体技术的在亚太社会的扩散、使用与效果，社会科学研究的本土化和国际化和等。先后获得国际传播学会 (ICA)、世界舆论研究学会（WAPOR）、美国舆论研究学会（AAPOR）、美国新闻与大众传播教育学会（AEJMC）等颁发的 7 项学术研究奖，发表论文 100 多篇，中文著作有《传播统计学》（合著）。曾任国际中华传播学会会长（2002－04 年）、《Journal of Computer-Mediated Communication》2002 专刊主编、美国 NSF 审稿人。现任《Human Communication Research》，《Communication Research》，《International Journal of Public Opinion Research》等 SSCI 期刊编委。

Let Social Sciences Ride on IT Bullet Train

Abstract:

Like sciences, the essence of social sciences is to process (human) information. Unlike sciences that have immediately benefited from the latest information technology, most of social science research has remained labor intensive in data collection and processing. We argue that digitalization and networking technologies have made it both necessary and feasible for social sciences to ride on the bullet train of IT. There are of course technical, economic, legal, ethical, and other issues including sampling, data sharing, privacy and confidentiality. However, these issues could be solved by sensible and user-friendly interface tools built jointly between IT people and social scientists. By helping social sciences to ride on IT’s bullet train, it in fact provides enormous opportunities for IT’s own growth.