Tim Berners-Lee Wendy Hall
James A. Hendler Kieron O'Hara 著
Nigel Shadbolt Daniel J. Weitzner

张 磊 江 勇 李涓子 译

A Framework for Web Science
万维科学的研究框架

清华大学出版社

A Framework for Web Science
万维科学的研究框架

Tim Berners-Lee Wendy Hall
James A. Hendler Kieron O'Hara 著
Nigel Shadbolt Daniel J. Weitzner

张 磊 江 勇 李涓子 译

清华大学出版社
北京

图书在版编目（CIP）数据

万维科学的研究框架/（美）李（Lee, T. B.）等著；张磊等译. —北京：清华大学出版社，2008.12
书名原文：A Framework for Web Science
ISBN 978-7-302-18899-5

Ⅰ. 万…　Ⅱ. ①李…②张…　Ⅲ. 万维网－研究　Ⅳ. TP393.4

中国版本图书馆 CIP 数据核字（2008）第 177560 号

责任编辑：龙啟铭
责任校对：徐俊伟
责任印制：李红英
出版发行：清华大学出版社　　**地　　址**：北京清华大学学研大厦 A 座
http://www.tup.com.cn　　**邮　　编**：100084
社　总　机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：北京市清华园胶印厂
经　　销：全国新华书店
开　　本：140×203　　**印　　张**：4.25　　**字　　数**：105 千字
版　　次：2008 年 12 月第 1 版　　**印　　次**：2008 年 12 月第 1 次印刷
印　　数：1～2000
定　　价：16.00 元

本书如存在文字不清、漏印、缺页、倒页、脱页等印装质量问题，请与清华大学出版社出版部联系调换。联系电话：010-62770177 转 3103　　产品编号：031856-01

序

信息技术革命深刻地改变着人们的学习、工作与生活方式，这其中互联网与万维网的发展更是在很短的时间内就显示出它们巨大的影响力。如今，借助于互联网与万维网，人们从事的活动非常的广泛与丰富，包括查找有用的信息、电子商务、电子政务、远程教育、网上购物、看新闻、听音乐、看电影、玩游戏、交朋友等等，可以说已经渗透到了日常的方方面面。不仅如此，世界上不管是发达国家还是发展中国家的人们都在使用互联网和万维网，可以说已经渗透到了世界的角角落落。人们都在享受着互联网和万维网带给我们的便捷。

但任何事物都是一把双刃剑，互联网与万维网也不例外，它给我们带来众多好处的同时，也会有一些负面的东西，比如上网成瘾、隐私侵犯、版权侵犯、网上欺诈、谣言传播、网上色情等等，而正是由于互联网和万维网本身的广泛性，使得这些负面影响也具有极大的广泛性和极强的危害性。这些势必影响到我们构筑一个和谐的社会。

对于具有如此大正面和潜在负面影响的事物，建立一门以其为核心研究对象的新学科，我认为是具有远见的，同时也是非常迫切的。很高兴看到由业界的几位领军人物提出建立万维科学（Web Science）这个学科，这样可以使更多各色各样相关领域的研究人员聚拢在这面旗帜之下，精诚协作，展开更深入和更有效的研究。

得知在万维科学方面具有强大实力的英国南安普敦大学与清华大学在深圳合作建立“清华——南安普敦网络科学深圳实验室”，以便在此领域展开高水平和深层次的合作研究，我认为这

是具有重大现实意义的一件好事，因为最新互联网权威统计显示中国网民总量已经在世界上排名第一，这说明互联网与万维网对于中国人的影响已经相当广泛了。我殷切期盼“清华——南安普敦网络科学深圳实验室”能够做出一些具有较高学术价值同时又利国利民的好研究。

“清华——南安普敦网络科学深圳实验室”翻译引进本书，我欣然为之写序，因为我想对于想进入该领域从事相关研究的人员，这是一本值得开始就好好看看的书，它为这个新学科指明了方向，为新进入该领域的研究人员提供了一个高层的全局图景。

万维科学的根本目的是要研究和理解万维网，以利用这些获得的认识进行趋利避害，通过技术或者社会的手段来使社会整体从万维网中受益。我想这自然是一个正确的方向，任何一个技术或学科的发展都应该走这样的方向。有了正确的方向，我相信这个学科一定会得到蓬勃发展。

衷心希望万维网领域可以在万维科学的旗帜下发展得更快！更好！

译 者 序

万维网的发展出乎所有人最初的预料。它深刻影响着我们的社会、工作、生活和未来。特别是在中国网民数量超过美国成为世界第一的时候，能够及时的把万维网之父和世界上其他顶尖科学家的这本著述翻译出来，我们深感幸运。

本书提出了一门新的科学——万维科学并规范其研究框架。由于其研究对象是万维网，我们将这门科学翻译为万维科学。但是清华大学和南安普顿大学的联合实验室使用的是“网络科学实验室”的名字，主要是因为实验室的大量研究内容特别是和产业应用相关的技术研究不可避免的要与底层的网络技术和硬件支持密切相关。

“同一个世界，同一个万维网”，这是2008年世界WWW大会在北京召开时候的口号。本书所探讨的万维网的科学、工程和社会人文经济的各个层面的问题，对于中国万维网的学者和从业者来说，同样的重要。特别是中国年轻一代，他们每天有大量时间生活在网络上、说着火星文、从事着人肉搜索并试图发出自己声音。而我们对此现象背后的原因极其可能产生的影响知之甚少。万维科学提供了一种思路和指引，使得我们可以用科学的方法去理解万维网和上面的使用者行为，并且确保社会利益和新的万维网经济模式可以从中产生。

这本书不是结束，而是一种开始。我们希望借由此书中文版的出版，万维科学从此可以引起更多人的关注、更多学者的研究讨论和更多万维网从业者的参与，并且一起创造万维网明天的繁荣和和谐。

本书第1章、第6章和第7章由江勇翻译，第2章和第3章

由李涓子翻译，第4章和第5章由张磊翻译。

最后，谨就我们由于学术功力不逮而可能出现的翻译遗误，向读者致歉，祈不吝指正。

译者

摘　要

本书建立了一系列用于万维网和类万维网信息结构分析与构造的措施，提出了一系列丰富的研究问题，同时进行了子学科划分，突出了万维网的多面特性及对之进行研究与开发的多学科特性。 这些问题和措施把作为分布式信息系统科学的万维科学（Web Science）提到了研究日程上。作为理解万维网的途径及专注关键信息通信和信息表示需求方面万维网开发的途径，我们需要万维科学。本文对诸如语义万维网（Semantic Web）、万维网服务（Web Service）及 P2P 之类的核心工程问题进行了综述。本文还讨论了用于发现万维网拓扑或其类图结构的分析方法。最后，由于万维网是一项与社会紧密融合的技术，因此也回顾了关于万维网使用与治理的各种问题与需求。

目　录

第1章 简 介

万维网技术的历史不算很长，但它的增长速度以及它对其植根的社会所产生的影响却令人吃惊。万维网技术最开始是为支持高能物理研究对信息的需要而诞生的。随后以迅不可挡之势扩散到其他科学领域、通用学术研究、商业、娱乐、政治以及几乎任何以通讯为目的的地方[142，143]。科学研究成果以及科学研究赖以进行的数据，不受打印和物理分发的约束，能够快速共享。链接使得科研工作处于丰富的上下文环境中。同时，革新拓宽了通信的可能性。博客（Weblog）和维基（wiki）允许人们即时会话，多媒体及交互性的潜力因此而变得巨大。

然而，无论是万维网还是现实世界都不是静态的。出于对来自科学、商业、公共与政治各方压力的回应，万维网在不停进化。例如，电子化科研（e-science）的成长使万维网需要集成大量不同的异构的数据，电子政务（e-government）和电子商务（e-commerce）也要求更有效地使用信息[34]。我们需要理解这些进化和发展的力量。如果没有这样一种理解，通过促进信息通信和信息表示的更多可能而为万维网增添价值的机会就可能丢失。但发展并不是全部。虽然万维网具有多面特性和可扩展性，但万维网还是基于一整套架构原则的，这些架构原则需要得到尊重。此外，万维网是一项社会性的技术，它在快速地增长，因此需要得到一个不断扩大的用户群的信任——可信任度、个人对信息的控制以及对他人的权利和偏好的尊重都是万维网非常重要的方面。这些方面都需要在万维网发展过程中得到理解和维护。

能够帮助识别什么是需要固定下来的以及哪里进行改变又是有利的一个研究日程是非常重要的。这正是万维科学的目标，这

个目标就是要勾画出分布式信息结构是如何服务于这些科学、信息表示及信息通信的需求的，并产生设计或者设计原则用以治理这种结构。我们认为这门关于分布式信息结构的科学对于理解人、代理（agent）、数据库、组织机构和其他角色和资源间随意未规划的信息链接是如何满足诸如科学研究信息化和电子商务之类的重要驱动力量的信息需要的是很重要的。一个高度分散系统如何能够具备所设计的性能表现是万维科学的关键问题。

万维科学（Web Science）是一个刻意含糊的词汇。物理学科是一门分析学科，它旨在找到产生或解释所观察到的现象的规律。计算机科学主要是（虽然不完全是）构造性的，因为形式化方法或算法的建立都是为了支持特定的需要的行为。万维科学则必须把这两种模式结合起来。需要研究和理解万维网，同时还需要用工程的方法构建万维网。微观上，万维网只是由人工语言和协议构建的基础设施，它是一项工程。但支配万维网的链接哲学及它在通信中的使用，使得万维网在宏观上表现出特性（它们中，有些是我们需要的，将之纳入工程，另外一些是我们所不愿留下的，可能的话就从工程中把它剔除）。当然，万维网在通信中的使用只是习俗和法律制约下人类交互更广阔系统中的一部分。万维网技术与人类社会在不同层面相互作用都意味着交叉学科特性是万维科学的固有要求。

这样一个能够用社会和科学的有用方法促使万维网发展的交叉学科的研究日程还没有提出来，我们需要创建它。最终，于2005年9月在英国伦敦召集了一个“万维科学研讨会”（Web Science Workshop）（在“感谢”中列出了对该研讨会有贡献的人员）。该研讨会中讨论了一些问题，包括：

- 万维网新趋势。
- 理解和指导万维网发展的挑战。
- 组织相关研究用以挖掘其他方面诸如普适计算、移动性、

新媒体和越来越多的在线数据所带来的机会。

- 确保诸如隐私的重要社会属性得到尊重。
- 识别和保护万维网体验中本质不变的东西。

本文源自该“万维科学研讨会”，并试图对该研讨会中讨论的东西作总结、延伸和评论。与会人员一致认为需要有交叉学科的措施，包括：计算机科学与工程、物理和数学科学、社会科学和策略制定。因此，**万维科学**不仅仅是关于不同微观和宏观层次上对万维网建模、分析和理解的方法。它也是关乎工程构建协议、提供基础设施和确保这些基础设施及其所附与的社会是相适应的。万维科学必须协调工程与社会，协调策略与技术约束及技术可能性，协调分析与构造。因此，它注定是交叉学科性质的。本文的章节组织也反映了这一点。

要发展万维网也涉及确认什么因素影响万维网体验，确保这些因素继续发挥作用。支撑万维网的基本架构决策的例子包括：404 错误（该错误表示链接资源失败但不产生灾难性后果）；统一资源定位符 URI 的使用；对现存互联网基础设施（例如，域名系统）的充分利用，并将之作为万维网构建的平台。标准也至关重要，W3C 负责创建和推荐标准，同时维护关键各方意见的一致，这表明工程需要和社会磋商过程携手同行。

本书第 2 章详细回顾那些基本的科学和架构原则。通过使用“进化”隐喻来帮助我们把万维网看作是一种既成的生态系统和一个具有政策与规则等一般社会需求的社会。关联相关措施、覆盖不同方法、变换不同的时空尺度和跨越一个宽大的领域范围进行建模非常具有挑战性。

第 3 章关注万维网工程的一些问题和如何推动诸如网格或服务之类的新技术以及如何让这些新技术推动万维网的问题。该章要讨论的可能最重要最具潜力的发展之一是语义万维网（Semantic Web）。万维网通常被刻画为相互链接文档的网络，这

些文档大多是用于人阅读的，因此机器可读性就需要自然语言处理的启发环境。然而，语义万维网，这种对万维网进行扩展和增值的想法，试图挖掘在链接的关系数据上进行逻辑判定的可能性以便能够自动化地处理更多的信息。开发能够支持查询、推理、修整数据模型、可视化和建模的语言和形式化方法，这样的研发工作已经有一些时日了。

繁荣语义万维网需要和万维网一样的分散哲学。一大挑战是确保局部一致的单个数据系统能够联合在一起，而不用试图做全局一致这样不可能完成的事情。此外，不用假设任何集中的或基础的形式化方法，而是通过具有诸如规则和逻辑等对比属性的一些形式化方法来使用常见的一套符号（即 URI），也是很重要的。第三个问题是关于把数据放在一起来利用合并以及任意重用的可行性。当前大多数数据都在单独的数据库中而没有发布（相比一般情况下文档向广大用户开放的 WWW 而言）。

第 4 章考察一些试图使用能够反馈到工程的方式来分析万维网的尝试。例如，以数学的方法对万维网建模可以使搜索和信息检索跟上万维网增长的步伐，特别是在联系一些诸如自然语言处理、网络分析和进程模型等重要领域的情况下。理解其出现的结构和宏观拓扑将帮助发现万维网所遵循的连接性和扩展规律。

众所周知，万维网的价值依赖于社会对它的使用，以及其在不破坏其他有价值的交互类型的前提下服务于通信需求的能力。这需要理解这些需求，理解其与其他社会结构的关系，理解其与技术发展间的双向作用。诸如这些社会问题在第 5 章讨论，包括处理象征意义的哲学问题；诸如推理方法等的逻辑问题，信任建立与维护的社会问题，以及通过万维网上人们的活动勾画社区。

一些社会与万维网技术间的相互作用已经广泛存在并需要管理和表达偏好的策略。例如，语义万维网显然激励着对于发布和共享数据资源极重要的社团及个性文化，这转而需要处理访问控

制、隐私、身份和知识产权的政策（以及能够向不同的用户群表达政策规则的接口和系统）。诸如这些政策、治理和政治问题在第 6 章讨论。

第 7 章提供了一个简单的小结，总结了万维科学的状况，并进一步扼要概括了文中已述的展望。

第 2 章　万维网与万维科学

我们可以把“万维科学”解释为是关于万维网的科学。尽管这种解释显而易见，我们还是需要仔细分析“万维科学”这个词语，描绘能够使万维网像一个分布式信息系统一样更有效工作的功能组件。在本章 2.1 节中，我们将回顾万维网体系结构的基本原理，阐述万维网上不断增强的信息共享和可信行为的社会价值。2.2 节将介绍一些在万维网上进行科学研究的新方法论。

2.1　万维网体系结构

万维网体系结构使用简单技术有效地连接各类信息，使信息空间高度灵活、可用，更重要的是信息空间的可扩展性。目前，万维网已经发展成为一个巨大的信息共享平台，取得了丰硕的成果。人们希望它能够更进一步发展，包括发展更多的语言、更多的媒体形式和更多的活动，拥有更多的信息，同时提供对万维网上信息进行查询的更好的工具和方法。这一节中，我们将简单介绍万维网体系结构的基本原理，其中参考了文献[155]，更细致的描述可以查阅该文献。

万维网是一个巨大的信息空间，在该空间中，资源由统一资源标识符（Uniform Resource Identifiers，URI）标识。万维网有支持主体间信息交互的协议，有用来表示万维网上信息的格式。这些是万维网的基本成分。万维网交互的可用性和有效性依赖于这些协议和信息的表示格式，依赖于许多原理，其中一些沿用万维网发展的最初概念，而一些是从万维网使用中学习到的经验。

为了更好地共享万维网上的信息、在万维网上进行推理、进