



链滴

# 外文资料翻译——The Semantic Web (语义网)

作者: [aopstudio](#)

原文链接: <https://ld246.com/article/1583419759476>

来源网站: 链滴

许可协议: [署名-相同方式共享 4.0 国际 \(CC BY-SA 4.0\)](#)

译者注：这篇文章是万维网之父蒂姆·伯纳斯·李于2001年发表于《科学美国人》杂志上的文章。在文中作者阐述了一种新的网络构想，即语义网，它不仅仅能让人阅读网页中的内容，还能让机器理解其中的含义。目前网络上暂时没有找到这篇文章的完整中文版。一些语句因为两种语言表达方式的差异能不太通顺，请大家多多包涵。

## 语义网

一种对计算机发展有深远意义的新的Web内容形式，可能产生新的知识革命

by TIM BERNERS-LEE, JAMES HENDLER and ORA LASSILAo;p

当电话响起时，音箱里正在播放披头士的《We Can Work It Out》。皮特接起电话时，电话通过向有拥有音量控制功能的本地设备发送了消息以调低音量。他的妹妹露西从诊所向他打来了电话，说道“妈妈需要去医院看专科医生，之后需要接受一系列的治疗。每隔两周，我要让我的代理(agent)（**者注：agent这词在后面多次用到，感觉翻译成代理并不太贴近，但也暂时想不到更好的表述**）去进预约。”皮特立即同意承担司机的工作。在诊所，露西通过她的手持浏览器对语义网代理下达指令。代理很快从医生的代理那里得到了关于妈妈处方治疗的信息，之后查询了一些提供相关治疗的医院的列表，并检查了那些能用妈妈的医保，同时在她家方圆 20英里 范围内，同时信用评级非常好的医院。之它开始通过查询可用的预约时间段（由不同医院的代理在它们的网站上提供）和皮特以及露西的空闲程表来确定预约的时间。（斜体字代表它们的语义或意思由代理通过语义网定义）

过了几分钟代理给他们展示了计划。皮特不喜欢这个计划，因为选择的大学附属医院需要从妈妈家开一路穿过城镇才能到达，而他会在高峰时间段开车回来。于是他让他自己的代理重新检索，并对位置时间信息做了更严格的限定。露西的代理在本任务中完全信任皮特的代理，并自动给它提供了刚才查出的信息的访问权限和快捷方式以协助它。立刻一个新的计划展示了出来：一个更近的门诊部和更早时间段——但是有两个注意事项。第一个是皮特需要重新安排一些不太重要的见面会议。皮特检查了些具体是什么会议，认为这没有关系。另一个注意事项是保险公司没有把这个门诊部列入提供治疗的单。“服务类型和保险计划状态通过其他方式得到安全验证。”代理向他保证。

露西在皮特嘟囔着“不用告诉我细节”的同一时刻表示同意，于是一切都安排好了。（当然，皮特无拒绝这些细节，那天晚上晚些时候，他的代理又向他解释了它是如何找到那个服务提供者的，尽管它在合适的名单上。）

## 表达语义

皮特和露西能够使用他们的代理去完成这所有的任务，要感谢的不是今天的万维网，而是将来它将会生的语义网。现在大多数的网页内容都是被设计给人来阅读的，而不能让计算机来有意义地操纵它们计算机可以熟练地分析网页以用来布局和日常处理，比如这里是标题，那里有一个指向其他页面的链接。但是通常计算机并没有可靠的方法来处理语义：比如这是哈特曼和施特劳斯诊所的主页，这个链接向哈曼医生的工作履历。

语义网将会为网页的的有意义内容带来结构，创造一个软件代理能够在页面之间漫游以方便地为用户行复杂任务的环境。这样当一个代理来到门诊部网页时，它将不仅仅知道页面上有“treatment, medicine, physical, therapy”这几个关键字（如今可能会被编码），同时也知道哈曼医生在星期一、星期二、星期五的时候在门诊部工作，脚本采用yyyy-mm-dd格式表示日期范围并返回预约时间。它将“认到”这一切，同时并不需要像《星球大战》的C-3PO和《2001太空漫游》的哈尔一样的人工智能。而代之的是门诊部管理员（从来没有学过Comp Sci 101课程）借助治疗协会网站上给出的资源并使现成的软件来编写语义网网页，即可将这些语义编码进网页中。

语义网并不是一个独立的网络，而是对现有网络的扩展，语义网中的信息被更好赋予含义，从而使得算机和人类能够协同工作。将语义网植入现有网络结构的第一步已经在进行。在不远的将来，随着机

能够更好地处理和“理解”它们目前显示的数据，这些发展将带来重大的新功能。

万维网的本质属性是它的普遍性。超文本链接的力量在于“任何事物都可以链接到任何事物”。Web 术因此不得区分潦草的草稿和优美的展现，无论是对于商业和学术信息，或在文化、语言、媒体等。息在许多的角度变化。其中之一就是主要为人类而产生的信息和主要为机器而产生的信息之间的区别在一端，我们拥有从五秒钟电视广告到诗歌的一切。在另一端我们有数据库、程序和传感器的输出。目前位置，Web在作为一种人类提供文档的媒介这方面发展得更为迅速，而不是作为一种可以自动处的数据和信息的媒介。语义网的目标是弥补这个不足。

和因特网一样，语义网将尽可能地去中心化。这样类似网络的系统在各个层面都会产生令人兴奋的事，从大公司到个人用户，并在将来提供难以预言的好处。去中心化需要妥协：一方面网络不得不抛弃有连接完全一致的理想，带来了臭名昭著的信息“Error 404: Not Found”；但同时允许了网络无制的指数增长。

为了语义网能发挥作用，计算机必须能够访问结构化的信息集合和可以用来进行自动推理的推理规则。早在网络发展以前，人工智能方面的研究者就已经对这方面的系统进行了研究。这个技术通常被叫知识表示，它目前的状态和网络出现以前的超文本相当：很明显是个不错的想法，同时有一些非常棒演示存在，但还没有改变世界。它包含了重要应用的种子，但要充分发挥它的潜力，它必须连接到一单一的全球系统中。传统的知识展示系统通常是中心化的，要求每个人对“parent”或“vehicle”样的概念定义完全相同。但是中心控制是令人窒息的，并且增加了系统的大小和规模。但是中央控制在窒息，增加这样一个系统的规模和范围它将很快变得难以管理。此外，这些系统通常会小心地限制以被问到的问题，以便计算机能够可靠地回答——或者说完全可以回答。这个问题让人想起数学中的德尔定理：任何复杂到有用的系统也包含无法回答的问题，就像复杂的基本悖论“这句话是错的”的杂版本一样。为了避免这样的问题，传统的知识表示系统一般都有自己的一套狭隘而特殊的规则来对据进行推理。例如，一个族谱系统，作用于一个家谱数据库，可能包括“叔叔的妻子是姑姑”的规则即使数据可以从一个系统传输到另一个系统，以完全不同的形式存在的规则，通常也不能进行推理。

相比之下，语义网的研究者则认为，悖论和无法回答的问题是实现多功能性必须付出的代价。我们使则的语言具有所需的表达能力，以允许Web根据需要进行广泛的推理。这种理念类似于传统的Web在Web发展的早期，批评者指出，它永远不可能是一个组织良好的库；没有一个中心数据库和树结构就永远不可能找到所有的东西。他们是对的。但是这个系统的表现力使得大量的信息可用，搜索引擎十年前看起来相当不切实际)现在可以生成大量材料的非常完整的索引。因此，语义网的挑战在于提一种既能表达数据又能表达规则的语言，用于对数据进行推理，并允许将任何现有知识表示系统中的则导出到Web上。

将逻辑添加到Web中使用规则进行推理、选择行动方案 and 回答问题的方法是语义网社区目前面临的任。数学和工程决策的混合使这项任务复杂化。逻辑必须足够强大，能够描述对象的复杂属性，但不能大到让代理考虑一个悖论而被欺骗。幸运的是，我们想要表达的大部分信息都是按照“六角头螺栓是种机械螺栓”这句话来表达的，这句话很容易用现有的语言编写，而且有一些额外的词汇。

开发语义Web的两项重要技术已经就位：可扩展标记语言 (XML) 和资源描述框架 (RDF)。XML 许每个人创建自己的标记隐藏标签，如 `<zip code>` 或 `<alma mater>`，用于对网页或页面上的文本分进行注释。脚本或程序可以以复杂的方式使用这些标记，但是脚本编写器必须知道页面编写器使用个标记的目的。简言之，XML允许用户在文档中添加任意的结构，但对结构的含义只字未提 (参见Jon Bosak和Tim Bray的《XML和第二代Web》；Scientific American, 1999年5月)。

意义是用RDF来表达的，RDF把它编码成三元组，每个三元组就像一个基本句子的主语、动词和宾语这些三元组可以使用XML标记编写。在RDF中，文档断言特定的事物 (人、网页或其他事物) 具有具特定值 (另一个人、另一网页) 的属性 (如“是谁的姐妹”、“是什么网页的作者”)。事实证明，种结构是描述机器处理的绝大多数数据的自然方式。主题和对象都由通用资源标识符 (Universal Resource Identifier, URI) 标识，就像在网页的链接中使用的那样 (URL, 统一资源定位符, 是最常见的RI类型)。这些动词也由URI标识，URI允许任何人定义一个新概念，一个新动词，只要在Web上的个地方为它定义一个URI。

当使用同一个术语来表示一些不同的事物时，人类语言蓬勃发展，但自动化却不是。想象一下，我雇

一个小丑信使服务，在我的顾客生日时给他们送气球。不幸的是，该服务将我的数据库中的地址传输到它的数据库中，不知道我的数据库中的“地址”指的是账单发送的地方，而且其中许多是邮局的邮箱。我雇来的小丑最后招待了一些邮政工人，不一定是坏事，但肯定不是预期的效果。为每个特定的概念用不同的URI可以解决这个问题。作为邮寄地址的地址可以与作为街道地址的地址区分开来，并且两者都可以与作为语音的地址区分开来。

RDF的三元组形成了关于相关事物的信息网。因为RDF使用URI在文档中对这些信息进行编码，URI概念不仅仅是文档中的单词，而是绑定到每个人都可以在Web上找到的唯一定义。例如，假设我们可以访问各种数据库，其中包含有关人员的信息，包括他们的地址。如果我们想找到居住在特定邮政编码的人，我们需要知道每个数据库中哪些字段表示姓名，哪些表示邮政编码。RDF可以指定“(数据库中的字段5) (是一个类型的字段) (邮政编码)”，为每个术语使用URI而不是短语。

## 本体论

当然，这并不是故事的结束，因为两个数据库可能对同一种概念使用两种不同的定义，比如邮政编码。一个程序想要比较或合并两个数据库中的信息，就必须知道这两个术语的意思是相同的。理想情况下程序必须有办法发现它遇到的任何数据库的共同含义。

语义网的第三个基本组件（称为本体论的信息集合）提供了这个问题的解决方案。在哲学中，本体论关于存在的性质、存在的事物的类型的理论；本体论作为一门学科来研究这些理论。人工智能和网络研究人员已经为他们自己的行话选择了这个术语，对他们来说，本体是一个文档或文件，它正式定义了语之间的关系。最典型的Web本体有一个分类法和一组推理规则。

分类法定义对象的类和它们之间的关系。例如，可以将地址定义为位置类型，并且可以将城市代码定义为仅应用于位置，等等。类、子类和实体之间的关系是一个非常强大的用于Web的工具。我们可以将属性分配给类并允许子类继承这些属性来表示实体之间的大量关系。如果城市代码必须是城市类型并且城市通常有网站，那么即使没有数据库将城市代码直接链接到网站，我们也可以讨论与城市代码联的网站。

本体论中的推理规则提供了进一步的动力。一个本体可以表达规则“如果一个城市代码与一个州代码关联，并且一个地址使用了该城市代码，那么该地址就有了相关的州代码。”然后一个程序可以很容易地推断，例如康奈尔大学地址位于伊萨卡，必须位于美国的纽约州，因此应按照美国标准格式化。计算机并不能真正“理解”这些信息，但它现在可以更有效地操作这些术语，其方式对人类用户是有用和意义的。

随着Web上的本体页面，术语（和其他）问题的解决方案开始出现。Web页面上使用的术语或XML码的含义可以通过从页面到本体的指针来定义。当然，如果我指向一个将地址定义为包含邮政编码的体，而您指向一个使用邮政编码的本体，则会出现与以前相同的问题。如果本体（或其他Web服务）供了等价关系，这种混淆就可以解决：我们的一个或两个本体可能包含我的邮政编码与您的邮政编码价的信息。

当两个数据库指向不同的地址定义时，我们发送小丑来招待我的客户的方案部分得到解决。程序对不同的地址概念使用不同的URI，这样就不会混淆它们，事实上还需要发现这些概念是相关的。然后，程可以使用一个服务，该服务获取一个邮政地址列表（在第一个本体中定义），并通过识别和移除邮箱和其他不合适的地址将其转换为一个物理地址列表（第二个本体）。本体论所提供的结构和语义使业家更容易提供这种服务，并使其使用完全透明。

本体论可以在许多方面增强Web的功能。它们可以简单地用于提高Web搜索的准确性——搜索程序查找那些引用了精确概念的页面，而不是使用模糊关键字的所有页面。更高级的应用程序将使用本体页面上的信息与关联的知识结构和推理规则相关联。在<http://www.cs.umd.edu/~hendler> 提供了一标记用于此用途的页面示例。如果你将网页浏览器发送到该页面，你会看到一个名为“詹姆斯A.亨德博士”的普通网页。作为人类，你可以很容易地找到一个简短的传记笔记的链接，并在那里阅读到亨勒从布朗大学获得博士学位。然而，一个试图找到这些信息的计算机程序，要想猜测这些信息可能在记中，并理解传记中使用的英语，就必须非常复杂。

对于计算机，该页面链接到定义计算机科学部门信息的本体页面。例如，教授在大学工作，他们一般有博士学位。页面上的进一步标记（不由典型的Web浏览器显示）使用本体论的概念来指定亨德勒是URI

<http://www.brown.edu>（布朗大学的网页）中描述的实体获得博士学位的。计算机还可以发现亨德是某个特定研究项目的成员，拥有特定的电子邮件地址，等等。所有这些信息都很容易被计算机处理并且可以用来回答目前需要人类筛选搜索引擎搜索到的各种页面内容的查询（比如亨德勒博士在哪里得学位）。

此外，这种标记使开发能够解决复杂问题的程序变得更加容易，这些问题的答案不在单个网页上。假你想找到去年在贸易会议上认识的库克女士。你不记得她的名字，但你记得她为你的一个客户工作，的儿子是你母校的学生。一个智能搜索程序可以筛选所有名字为“库克”的人的页面（避开所有与厨、烹饪、库克群岛等相关的页面），找到客户名单上提到为一家公司工作的人，并跟踪他们孩子的网链接一查看是否孩子的学校符合条件。

## 代理

当人们创建许多程序来收集来自不同来源的Web内容、处理信息并与其他程序交换结果时，语义网的正力量就会实现。随着更多的机器可读Web内容和自动化服务（包括其他代理）的可用，这些软件代的有效性将成倍增加。语义网促进了这种协同：即使没有明确设计为协同工作的代理，当数据带有语时，也可以在它们之间传输数据。

代理功能的一个重要方面将是用语义网的统一语言（表达使用规则和信息（如由本体论指定的规则和息）进行的逻辑推理的语言）编写的“证明”的交换。例如，假设库克女士的联系方式是通过一个在服务找到的，让你大吃一惊的是，她把她的地点定在了约翰内斯堡。当然，您需要检查这个问题，所您的计算机要求服务提供答案的证明，并通过将其内部推理转换为语义网的统一语言迅速提供答案。电脑里的一个推理机很容易验证这位库克女士确实与你要找的人相符，如果你还有疑问，它可以向你示相关的网页。尽管它们还远远没有深入挖掘语义网的潜力，但一些程序已经可以使用统一语言的当初步版本来通过这种方式交换线索。

另一个重要的特性是数字签名，它是加密的数据块，计算机和代理可以使用它来验证附加的信息是否特定的可信来源提供。你要确定发送给你的会计程序的一份声明，说你欠一家在线零售商的钱，并不隔壁精通电脑的青少年伪造的。代理应该对他们在语义Web上阅读的断言持怀疑态度，直到他们检查信息源。（我们希望更多的人能在网络上学习到这一点！）

许多没有语义的自动化的基于Web的服务已经存在，但其他程序（如代理）无法找到将执行特定功能服务。此过程称为服务发现，只有当有一种通用语言以一种让其他代理“理解”所提供的功能和如何用它的方式来描述服务时，此过程才能发生。例如，服务和代理可以通过将这些描述存放在类似于黄的目录中来宣传其功能。

目前，一些低级别的服务发现方案是可用的，例如专注于连接不同类型设备的微软通用即插即用（Universal Plug and Play）和旨在连接服务的Sun公司的Jini。然而，这些方案在结构或句法层面上有问题并在很大程度上依赖于预定功能描述集的标准化。标准化只能走到目前为止，因为我们无法预测未来有可能的需求。

相比之下，语义网则更加灵活。消费者和生产者代理可以通过交换本体论来达成共识，本体论提供了论所需的词汇。当发现新的本体时，代理甚至可以“引导”新的推理能力。语义还使得利用只与请求分匹配的服务变得更容易。

一个典型的过程将包括创建一个“价值链”，其中信息的子组件从一个代理传递到另一个代理，每个理“增值”，以构建最终用户所要求的最终产品。毫无疑问，为了按需自动创建复杂的价值链，一些理除了语义网之外，还将开发人工智能技术。但语义网将提供基础和框架，使这些技术更为可行。

将所有这些特性组合在一起，将导致皮特和露西的代理拥有在本文开始的场景中所展示的能力。他们代理将以分解的方式将任务委托给其他服务和通过服务广告发现的代理。例如，它们可以使用受信任

服务获取提供者列表，并确定哪些提供者正在为指定的保险计划和治疗过程制定计划。提供者的列表由另一个搜索服务提供，等等。这些活动形成了一系列的链，其中从原本大量分布在网络上的数据（这种形式下几乎毫无价值），逐渐减少到对皮特和露西来说很有价值的少量数据，这是一个符合他们程安排和其他要求的预约计划。

下一步，语义网将突破虚拟领域，扩展到我们的物理世界。URI可以指向任何东西，包括物理实体，意味着我们可以使用RDF语言来描述手机和电视等设备。这样的设备可以像软件代理一样宣传他们的能，他们可以做什么以及如何控制他们。这种语义方法比通用即插即用（Universal Plug-and-Play）等低级方案灵活得多，因此打开了一个充满令人兴奋的可能性的世界。

例如，今天所谓的家庭自动化需要精心配置才能使设备协同工作。对设备兼容性和功能的语义描述将我们在最少的人工干预下实现这种自动化。一个小例子发生在皮特接电话时，音响的声音被关了。他必为每一个特定的设备编程，而是可以一次编程从而一劳永逸地实现这样一个功能，这个功能覆盖每个有音量控制的本地设备——电视、DVD播放机，甚至是他今天晚上下班回家的笔记本电脑上的媒体播放器。

在这方面已经采取了第一个具体步骤，当前正在制定一个描述设备功能（如屏幕大小）和用户偏好的准。该标准建立在RDF之上，称为复合能力/偏好配置文件（CC/PP）。最初，它将允许手机和其他非准的Web客户端描述它们的特性，以便可以随时为它们定制Web内容。稍后，当我们添加处理本体逻辑的语言的全部通用性时，设备可以自动寻找并使用服务和其他设备来添加信息或功能。不难想象的网络微波炉咨询冷冻食品制造商的网站，以获得最佳烹饪参数。

## 知识革命

语义网并不仅仅是我们迄今为止讨论过的执行单个任务的工具。如果设计得当，语义网可以帮助产生个人类社会中的知识革命。

人类目前陷入了一种永恒的矛盾中，即小团体独立行动的有效性与与更广泛的社区建立联系的需要间的矛盾。一个小团体可以快速有效地创新，但这会产生一种亚文化，其概念不为他人所理解。然而在一个庞大的团队中协调行动是非常缓慢的，需要大量的沟通。世界在这两个矛盾之间运行，并存在种趋势——从个人或小团体的想法开始，随着时间的推移，逐渐变成一种在大团体中通用的理解。

当需要更广泛的共同语言时，一个基本的过程是亚文化的结合。通常两个群体独立地发展出非常相似概念，描述他们之间的关系会带来很大的好处。就像一本芬兰语英语词典，或者一个权值和度量转换，尽管概念的共同性还没有导致术语的共同性，但描述不同术语之间的关系将带来交流和合作。

语义网通过简单的URI命名每个概念，让任何人都可以很轻松地来表达他们发明的新概念。统一的逻辑语言将使这些概念逐步连接到一个通用的网络中。这种网络将使软件代理对人类的知识和工作进行有意义的分析，为我们的生活、工作和学习提供一类新的工具。