语义网

——一种能让计算机理解的新型Web内容形式

孙雄勇
（中国科学院声学所, 北京 100080）
Email: xiongyongsun@163.com

目前的万维网其进化、扩大和完善的空间还很大，可以说万维网还没有走出婴儿期。为使万维网迈上一个新的台阶，从此摆脱幼稚，走向成熟和真正的智能化，10年前为我们发明因特网超文本系统的麻省理工学院万维网协会主席蒂姆·伯纳斯·李，现在又在致力于开发新一代的万维网（互联网），他为之取了一个直观的名称——“语义网”（the Semantic Web）。

1、什么是“语义网”？

语义网就是能够根据语义进行判断的网络。

目前在万维网中，网页仅仅是一个单调的内容显示，电脑只负责将一个网页链接到另一个网页，网络不能按照用户的要求自动搜寻和检索网页，直至找到所需要的内容。而语义网则是希望计算机能“看懂”网页的内容，使计算机成为“智能”的导航工具。当然语义网还并不仅仅能完成这个功能，它比这还要“聪明”得多。

简单地说，语义网是一种能理解人类语言的智能网络，它不但能够理解人类的语言，而且还可以使人与电脑之间的交流变得像人与人之间交流一样轻松。

语义网就好比一个巨型的大脑，它由数据库智能化程度极高，协调能力非常强大的各个部分组成，可以解决各种难题。在语义网上连接的每一部电脑，都能分享人类历史上所有科学、商业和艺术等知识。它不但能够理解词语和概念，而且还能够理解它们之间的逻辑关系。

在语义网中，网络不仅能够连接各个文件，而且还能够识别文件里所传递的信息，也就是说，它是一种聪明的网络，可以干人所从事的工作。例如：它可以让计算机辨认和识别“head”这个单词的意思是“头脑”还是“领导”；在读者看新闻时，它能轻松地分辨出哪句是标题、哪句是导语。

2、语义网与万维网的区别

目前我们所使用的万维网，实际上是一个存储和共享图象、文本的媒介，电脑所能看到的只是一堆文字或图象，对其内容无法进行识别。万维网中的信息，如果要让电脑进行处理的话，就必须首先将这些信息加工成计算机可以理解的原始信息后才能进行处理，这是相当麻烦的事情。而语义网的建立则将事情变得简单得多。

语义网是对万维网本质的变革，它的主要开发任务是使数据更加便于电脑进行处理和查找。其最终目标是让用户变成全能的上帝，对因特网上的海量资源达到几乎无所不知的程度，计算机可以在这些资源中找到你所需要的信息，从而将万维网中一个个现存的信息孤岛，发展成一个巨大的数据库。

语义网将使人类从搜索相关网页的繁重劳动中解放出来。因为网中的计算机能利用自己的智能软件，在搜索数以万计的网页时，通过“智能代理”从中筛选出相关的有用信息。而不像现在的万维网，只给你罗列出数以万计的无用搜索结果。

例如，在浏览新闻时，语义网将给每一篇新闻报道贴上标签，分门别类的详细描述哪句是作者、哪句是导语、哪句是标题。这样，如果你在搜索引擎里输入“老舍的作品”，你就可以轻松找到老舍的作品，而不是关于他的文章。

总之，语义网是一种更丰富多彩、更个性化的网络，你可以给予其高度信任，让它帮助你滤掉你所不喜欢的内容，使得网络更像是你自己的网络。

3、语义网的实现

语义网虽然是一种更加美好的网络，但实现起来却是一项复杂而浩大的工程。

要使语义网搜索更精确彻底，更容易判断信息的真假，从而达到实用的目标，首先需要制订标准，该标准允许用户给网络内容添加元数据（即解释详尽的标记），并能让用户精确地指出他们正在寻找什么；然后，还需要找到一种方法，以确保不同的程序都能分享不同网站的内容；最后，要求用户可以增加其他功能，如添加应用软件等。

语义网的实现是基于XML（可扩展标记语言eXtensible Markup Langauge）语言和资源描述框架（RDF）来完成的。XML是一种用于定义标记语言的工具，其内容包括XML声明、用以定义语言语法的DTD (document type declaration文档类型定义)、描述标记的详细说明以及文档本身。而文档本身又包含有标记和内容。RDF则用以表达网页的内容。

当然，要实现语义网并非仅有XML和RDF就行了。更主要的技术难题还在于要让电脑可以进行过多的“思考”和“推断”，而

面对纷繁复杂的问题，尤其是社会问题，人尚且难以决断，更何况计算机呢。因此，要真正实现实用的语义网还有很多工作要做。

4、**XML**和语义

XML的最突出的特点就是功能强大又易于使用，它使网页能够容纳更丰富的信息资源。其中元数据管理、语义透明性和自主主体都是XML所独有的概念，而XML对统一结构化语法和半结构化语法的承诺，将有助于把几乎不可能完成的事变成切实可行的。

那么在XML的基础上所讲的语义又是什么呢？虽然语义这个单词每个人对其定义的观点各有不同，但一般来说，我们可以将语义解释为构建在公用语法之上的系统中XML数据的一层规范。这就引出了下面标记了XML语义的概念（在下面三概念之间有一些重叠）：

元素类型名称、属性名称和某些情况下内容术语的解释；

用于使用有效文档引导事务的处理规则（也称作商业规则）；

一个文档中的结构化元素与另一个文档中的结构化元素之间的关系。

5、知识表示：

为使语义网工作，计算机必须能访问结构化的信息集合以及一套推理规则，据此进行自动推理，HNC在这方面应该大有用武之地。在Web被开发出之前很久，人工智能研究人员就已经研究过这样的系统。这个技术通常称为知识表现，和Web出现之前的超文本的境地类似：它的确是个好主意，也有一些非常好的范例，但是它还无法影响和改变世界。它蕴含了能产生重要应用的种子，但是要充分发挥其潜能，它必须和一个全球系统联系在一起。

传统的知识表现通常是集中化的， 要求每个人对于共同的概念，如“父母”和“汽车”，使用完全一样的定义。但是，集中化控制比较死板，而且这种系统的规模和范围增长过快，很快会变得难以管理。

除此之外，这些系统往往小心地对允许问的问题加以限制，这样计算机才能给与可靠的回答。问题就像数学中的哥德尔理论：任何足够复杂的系统如果是可用的，就必然存在不可解决的问题。 也就像那个最基本的悖论的复杂版本： “本句话是错误的。”为避免此类问题，传统的知识表达系统通常各自都有针对其数据作推理的一套有限和特殊的规则。例如，一个基于家庭数据库的家谱系统，可能包含规则“叔叔的妻子是婶婶”。即使数据可以由一个系统传到另一个系统，规则则不然，由于规则所处的环境完全不同了，它往往不能运用到另一个系统中了。

相反，语义网的研究者认为要获得多样性，必然会有自相矛盾的情况或无法回答的问题出现。描述规则的语言要尽量具有表达力，让Web能尽可能广泛地进行推理。这个思想和传统的Web相似：在Web开发的早期，恶意批评者指出它永远无法是一个组织良好的库； 没有集中的数据库和树状结构，人们无法确保找到任何东西。他们曾经是正确的。但是，系统的表现能力使我们能获得大量的信息，而搜索引擎（十年前看起来不切实际）现在能从中对许多材料产生出非常完整的索引。因此，摆在语义网面前的挑战是，提供一种语言，能同时表达数据以及根据数据进行推理的规则，并且允许任何现存的知识表现系统中的规则都能输出到Web上。HNC对自然语言语义网络的基本构成及其特性进行了一个总体性的描述。我们应该可以在利用HNC理论在web上实现某些功能，譬如智能检索（在以后的文章中将进行讨论）。

在Web上增加逻辑性—使用规则去推理、选择行为的步骤并回答问题的方法—是语义网组织面临的一个任务。这个任务涵盖了数学和工程化决策，使其更加复杂。逻辑必须强大到能够描述复杂的对象属性，但也不能太复杂，导致代理可能被一些悖论的问题问倒。幸运的是，通常我们大部分想表达的意思就像“六头螺钉是一种机器螺钉”这样的句子，稍加一些词汇表，用现在的语言就能将其表达出来。

6、语义网的优点

建立语义网的重要性在于，对信息含义的理解不再是只有依靠人才能完成的事情，计算机同样也可以完成这样的工作。

例如，我们看到网页上的天气预报，自然就会知道其中的含义，但计算机并不知道在那么多的数字中，哪一个数字代表温度，哪一个数字代表湿度。而语义网的意义就要在隐藏的编码中，指明哪个数字代表温度，哪个数字代表湿度，并且说明“温度”和“湿度”的含义。

语义网最大的好处是可以让计算机具有对网络空间所储存的数据，进行智能评估的能力。这样，计算机就可以像人脑一样“理解”信息的含义，完成“智能代理”的功能。使用语义网搜索引擎搜索的结果也将比万维网更为精确。

另外，由于大部分科技创新和突破，都是对已有知识的重新组合和更新，因此语义网也为新的科技创新提供了无尽的资源，它可以在很短的时间内，完成一个人甚至需要一辈子才能做出的组合结果。

蒂姆·伯纳斯·李曾说过：“完全可以想象，一旦这种技术被运用于世界上所有的数据表格，它将产生极大的社会效益。”

据称美国将于2005年推出语义网。我们有理由相信，语义网一定会给我们带来互联网的新时代。

作者简介：孙雄勇（1978--），男，湖南邵阳人，1978年生，2001年7月从清华大学中文系计算语言学专业毕业并获得学士学位。现为中科院声学所博士研究生，指导教师张全研究员。专业：自然语言理解，主要是HNC自然语言处理理论及相关技术研究。