语义 Web 的实现:概念标记与概念系统*)

黄映辉　李冠宇

(大连海事大学计算机科学与技术学院　大连 116026)

摘　要　语义 Web 是对 Web 的扩展。Web 是被格式标记的信息的集合,语义 Web 则是被概念标记的信息的集合,扩展的两项措施为信息采用概念标记和计算机内置概念系统。语义由直接语义和引申语义构成,前者为人脑中的观念,后者与语境有关。鉴于目前计算机的能力尚不能模拟语境,不得不"搁置引申语义"和"以概念近似观念",于是就有"语义=概念"。概念标记与概念系统是实现语义 Web 的两大支撑。概念标记就是用概念标记符对将要交由计算机处理的信息进行标记,其面临的主要难点有信息的切分、概念标记符的选用和标记过程的自动化。概念系统就是 Ontology,其主要功用是读出信息的直接语义并进行概念推理,后者才是语义 Web 的最突出特征。概念系统的研究目前主要针对规模庞大、持续性维护、分布式及其集成应用等问题。

关键词　语义 Web,语义,概念标记,概念系统

Constructing Semantic Web: Concept Markup and Concept System

HUANG Ying-hui　LI Guan-yu

(College of Computer Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract　The Semantic Web is an extension of the Web. The Web is a set of the information marked up with the format tag, whereas the Semantic Web is one with the concept tag, and the two means of the extension are to apply the concept markup for the information and to build the concept systems within computers. The linguistic meaning is composed of the distinct meaning and the connotational meaning, the former is the idea in human brain, but the latter is related to the context wherein language using. As the ability of computer presently not enough to simulate the context, it cannot help but to lay the connotational meaning aside and to think the concept nearly being the idea, so that the definition of the linguistic meaning equal to the concept is deduced. The concept markup and the concept system are the two major supports for constructing the Semantic Web. The concept markup is to markup the information to be inputted into computer for processing with the concept tag, its difficulties at present are how to partition the information, how to select the concept tag, and how to realize the automation of markup process. The concept system is just the Ontology, its functions are to understand the distinct meaning of information and to carry out the concept inference, the latter is just the remarkable character. Nowadays, the majority of researches about the concept system focus on the problems of the huge scale, the durative maintenance, the distributed architecture and its integration applications.

Keywords　Semantic Web, Linguistic meaning, Concept markup, Concept system

1　引言

语义 Web 是对 Web 的扩展,这是同为 Web 与语义 Web 首倡者的 T. Berners-Lee 的精辟论点[1]。它不仅阐释了语义 Web 的性质,更重要的是指明了实现语义 Web 的方向:扩展的基础是 Web,扩展的目标是"语义"。由此引出了语义 Web 研发人员需要彻底理清的三个问题:Web 是什么?语义 Web 是什么(其核心是"语义"为何)?扩展的措施有哪些?以便有效地指导语义 Web 的具体实现。

2　Web 是被格式标记的信息的集合

2.1　Web 的定义

1990 年发明的 Web 是 World Wide Web 的简称,其汉译为"万维网"。中文的"网"总给人以实体的感觉,很多人望文生义认为 Web 是一种物理网络,特别是与 Web 联系在一起的"服务器"和"客户机"两种角色计算机的真实存在更加深了这种认识。

由 T. Berners-Lee 担任主席的 W3C 组织给出了权威性定义:Web 是一个信息空间,其中的信息项(item)被看作是资源,由统一资源定位符 URI 所标识[2]。因为 URI 的全球惟一性,这个定义所强调的是信息的"易被寻到"性质,彰显了 Web 促进实现信息共享之目标。

Web 实际上是基于 Internet 的一种信息服务(与之并列的还有 FTP、E-mail、BBS 等)。信息服务的提供者和使用者均为网民,信息服务的中介角色则是信息服务商。信息的地址由 URI 标识,信息的表示用 HTML,信息的传输遵守 HTTP。

2.2　Web 并非没有语义

语义 Web 构想的提出原本就是基于对"Web 不具有语义"的批评,但它却欠缺严谨。语义是语言符号对人的意义。

*)国家自然科学基金资助项目(60672031),辽宁省自然科学基金资助项目(20072142)。**黄映辉**　教授,博士,CCF 高级会员,研究方向为智能信息处理;**李冠宇**　教授,CCF 高级会员,研究方向为智能信息处理。

Web完全是一个人造系统。难道人类会费力去构建一个对自己没有意义的东西吗？Web今日之所以风靡全球，就是因为每一位阅读网页的人都从中获得了意义。Web具有语义毋庸置疑。只是此语义非彼语义。

T. Berners-Lee在阐释语义Web概念时指出：语义就是机器可处理[3]。可见，语义Web所强调的"语义"是计算机所能理解的语义。这才是Web所不具备的。严谨的表述应该是：Web上的信息具有人脑理解的丰富语义，而不具有计算机理解的语义；语义Web上的信息不仅具有人脑理解的丰富语义（此时需要将信息显示出来以使人能够阅读），也具有计算机理解的丰富语义。Web信息直接供人阅读的；语义Web信息直接供计算机进行处理，只将查询与推理的最后结果显示给人们。

2.3 被格式标记的信息

对于Web最通俗的表述为：Web是网页的集合。网页是为人们直接阅读而设计的。如何将信息在计算机屏幕上以最有利于人的视觉接收的格式显示出来，成为构建Web的关键。作为SGML（标准通用标记语言）简化子集的HTML就担当此任，对进入Web信息空间的每一条信息都用某些显示格式标记符（format tag）进行了标记（marking up）。

3 语义Web是被概念标记的信息的集合

3.1 语义Web的"语义"

3.1.1 问题的提出

当下对语义Web的积极探索，大多是以T. Berners-Lee所提出的七层"语义Web体系结构"[4]为基础，讨论"如何实现语义"，而对"语义究竟为何"却关注甚少。可以说，"语义"是目前语义Web研究中认识最为模糊的概念。究其原因有二：(1)在源出领域就争论不休。"语义"本是语言学和语言哲学的研究对象，从1825年K. Reisig首创语义学（Semasiology）[5]以来的理论纷争，迄今也未就"什么是语义"达成共识；(2)引入计算机领域后欠缺深入的对比研究。T. Berners-Lee的著名论断"语义就是机器可处理"也只是陈述了"语义的性质"并未回答"语义是什么"。需要研究的句子、内容很多，例如：计算机领域的"语义"与语言学领域的"语义"是否等同？语言学领域的"语义"与人脑相关联，而计算机领域的"语义"则与机器相对应，二者的相似与差异何在？

3.1.2 语义定义的不同观点

"语义Web"译自英文Semantic Web。形容词性的semantic（语义的）衍生于名词semantics（语义学），后者又源自法文sémantique，该词为法国语言学家M. Bréal借用希腊语词根sēma（符号）于1894年所创[5]。Semantics（语义学）与上述的Semasiology（符号学/语义学）同根。令人十分意外的是：英文中竟然没有与中文"语义"一词相对应的专用名词。而以词组的方式表示之：*linguistic meaning*或*lingual meaning*。对此现象的猜想是：之所以未新创一个以sēma为词根的名词，完全是为了强调"语义就是meaning（意义）"而淡化其与"符号"的联想。

语义即语言的意义。何谓"意义"？按照当代哲学主流学派语言哲学的观点，语言是表示思想的符号系统。符号（symbol）的实质是其代表性，即符号代表着某个或某些在符号之外的对象[6]，符号的意义就在于此。语义就是语言符号所指的对象。这个"对象"究竟为何却是语言学家以及哲学家们长期争论的问题。

语言的意义（即语言所指对象），一方面与语言符号保持着较为稳定的对应关系，另一方面也随着语言使用具体环境的不同而变化。语义学（Semantics）研究语言在形式上的意义，语用学（Pragmatics）则研究语言在语境中的意义。浏览语言学领域各种"意义理论"[7]，可有这样的结论：迄今的争论与发展主要集中在语用学范畴，而在语义学层面上学者们近乎达成共识，那就是指称论与观念论。

指称论（Referential Theory）认为，语义（语言所指对象）是真实世界中的一个或一类客观事物即指称（referent）。第一个系统阐述指称论的是英国哲学家密尔（J. Mill，1773－1836），后又受到大科学家罗素（B. Russell，1872－1970）的推崇。指称论简单直观，易被普遍人所理解。但是，语词与现实的两两相连、整齐对应，难以解释许多语言现象，其自身也包含着不少矛盾[7]。

观念论（Ideational Theory）认为，语义（语言所指对象）是人们头脑中的观念（idea），是它将语言与客观事物联系起来。观念论的源出为英国经验主义哲学家洛克（J. Locke，1632－1704）的名著《人类理解论》。洛克的界定遭到了德国数理逻辑学家弗雷格（G. Frege，1848－1925）等学者的批评：头脑中的观念因人而异，同一词语就可能意义不同，人们如何交流？弗雷格将观念论修正为：观念"可以是很多心灵的共同性质"[8]，即给出了普遍性和持久性的约定。相比之下，观念论比指称论有较好的解释能力，成为被广泛接受的观点。

3.1.3 计算机领域的选择：搁置语境

语言存在的必要在于人们的使用。拘泥于语言形式本身的指称论和观念论固有其天生的缺陷，要深入全面地探讨"语义为何"只有进入到语境（context）中才行。通常将指称论或观念论给出的定义称为直接语义（distinct meaning），而与语境有关的附加称为引申语义（connotational meaning）[6]。

语义Web构建在计算机系统内。目前的计算机在辨识和表达人的情感、暗喻、委婉等方面尚在起步阶段。也就是说，计算机还不能理解语言使用中的语境，也无法在机器中构建出模拟的语境来。因此，不考虑语境的影响而接受基于观念论的语义定义即"语义＝人脑中的观念"不失为一种可行现实的选择。这样，既可实现语义Web的初级应用，也为计算机进一步获得引申语义奠定了基础。

3.1.4 语义＝计算机中的概念

观念（idea）是客观事物在人脑里留下的概括的形象（辞海，1989），其外显形式是多样化的。概念（concept）是观念的一种类型，其表达仅用词或词组，因而具有更多的形式化特征。显然，概念更易被计算机所接受。计算机是对人脑的模拟，移植上述关于语义的定义，就有"语义＝计算机中的概念"。

3.2 信息用概念标记符标记

语言的交际就是语义的传递过程，它是这样实际发生的：当某人获得某个语言符号（语音或文字）时，就在其头脑中唤起相应的观念，并由此联想到该观念所指向的一类客观事物。语义三角形（C. Ogden & I. Richards，1923）是阐释此过程的著名模型[9]。用计算机实现这一过程的情景则是：当计算机读取某个符号时，就在其存储中唤起相应的概念，并由此联想到该概念所指的那些信息内容。可见，符号、概念、信息内容三位一体是实现机器"读懂语义"之关键。图1是基于语义三角形而建立的人与机器读懂语义的比较模型。

如图1所示，人脑中的观念对应的符号是语音或文字，它

们与观念的存在形式并不相同。计算机中的概念的存在形式是词或词组，其所对应的符号与之同形同义，只是前者早也存储在计算机内部，后者将由外部输入。该符号的主要功用是标记与概念所联系的信息内容，所以将之称为“概念标记符(concept tag)”。与语音或文字直接进入人脑不同的是，概念标记符则与信息内容构成 XML 元素再输入到计算机。当计算机在其内部找到了与该概念标记符完全一致的概念时，就将该信息内容与这个概念相绑定，即认为这就是该信息的语义。计算机读出了语义。

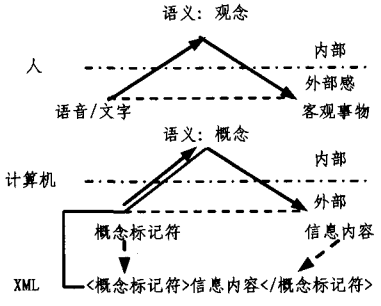

图 1　理解语义：人与计算机的比较

对信息内容用概念标记符进行标记(以下简称概念标记，concept markup)是计算机之间相互理解所交互信息的基础性工作。

3.3　语义 Web 是被概念标记的信息的集合

语义 Web 由 Web 扩展而来。基于上述讨论的结论，表 1 对比列出二者的主要特征，由此将语义 Web 定义为“被概念标记的信息的集合”是合适的。

表 1　Web 与语义 Web 的对比

类型	特征	性质	使用者
Web	格式标记	网页集合	人
语义 Web	概念标记	文档集合	机器

4　从 Web 到语义 Web 的扩展

4.1　两项措施

基于图 1 的分析，可概括语义 Web 的工作原理如图 2 所示。语义 Web 这样一个前沿性领域的所作所为可以用“概念”给出通俗化的解说：从实际领域抽象出概念系统；将概念系统输入到计算机；用概念系统中的概念作为概念标识符去标记所有将要输入到计算机中的信息；计算机读取已被概念标记过的信息，就会比照内置的概念系统进行识别，更重要的是可依据概念系统中所标示的概念之间的关系，从一个概念联系到另外的概念及其属性(即概念推理)，从而给出这些信息的意义来。所谓计算机读懂的信息语义，就是该信息能“带出来”的那些概念及其属性。

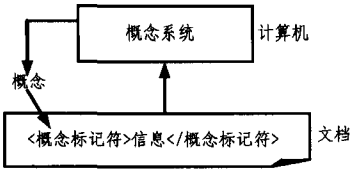

图 2　语义 Web 的工作原理

显然，支撑语义 Web 有两大基础性工作：计算机读取的信息全部用概念标记；计算机内部置放概念系统。

4.2　信息采用概念标记

正如第 3.2 节所作的定义，概念标记就是用概念标记符对将要交由计算机处理的信息进行标记。XML 是具体的实施技术，只是所采用的自定义标记符必须是领域共同认可的概念标记符。原理简单清晰，但操作起来尚有若干难点亟待克服：

(1)信息内容的切分。信息文档可以按段落、句子，甚至词或词组进行标记。切分得越细，计算机从该信息文档中可读出的语义越多，其不利之处则是：一方面标记的工作量加大，另一方面需采用更多的概念标记符，即要求在计算机内构建更庞大更复杂的概念系统。如何切分也是一个问题，对同一篇信息文档不同的人可能有不同的切分方案，计算机读出的语义也就有所差异。

(2)概念标记符的选用。选用的概念标记符的数量由信息文档的切分方案来决定。至于选用哪个概念标记符就受到两方面的约束：标记主体对所标记信息内容的真实意义的准确把握；计算机内置的概念系统是否能提供所需要的概念。

(3)标记过程的自动化。语义 Web 是被概念标记的信息的集合，海量的信息文档用人工标记是不可想象的。关于半自动/自动标记的方法与工具已有不少研究成果[10]，但是距离支持语义 Web 构建的实用目标尚有很长的路要走。

4.3　计算机内置概念系统

需要在计算机内置的概念系统就是始于 T. Gruber 发表于 1993 年的那篇著名论文[11]，而今在学术界热烈讨论的 Ontology。两种称呼同义，概念系统基于内部组成，Ontology 则强调其重要性：它是实现信息的机器自动处理的本体或根本。

4.3.1　概念系统的实际作用

类似于人脑中的知识网络，计算机中的概念系统主要由概念、概念的属性、概念之间相互关系三者组成，是一种层次网络。概念系统在语义 Web 中所起的作用主要有二：

(1)读出语义。前已定义，概念系统中的概念就是语义。当用概念作为标识符去标识信息时，就给该信息绑定了某个语义。当这个由信息及其绑定语义组成的 XML 元素被输入到计算机时，基于内置的概念系统计算机就会认定：该信息的语义就是这个概念。这一过程类似于人给一些密闭容器贴标签：先写上里面装的何物(概念标识)，以便下次见到时能认识(概念系统)。能读出信息的直接语义，仅相当于幼儿识字的水平。概念系统更重要的功能是概念推理，这才是语义 Web 优于 Web 的最根本一步。

(2)概念推理。作为层次网络的概念系统具有通达性(connectedness)，即从某个概念(属性)处出发就可达到与该概念相联接的其他概念(属性)。概念推理(concept inference)就是从一个概念(属性)导出若干相关概念(属性)。也就是说，使用概念系统计算机可以读出信息的多种语义来。

4.3.2　概念系统的构建难点

对 Ontology 的讨论十余年经久不衰，说明概念系统的构建确实有些难于解决的问题。

(1)规模庞大。计算机中的概念系统的本源一定是人脑中的概念系统。咋然一听，似乎只是一个输入问题。然而事实的真相却是：人脑中的概念及概念系统的总体特征是模糊性与多义性，而输入到计算机中的表现概念与概念关系的词或词组一定是明确的与单义的。为减弱模糊性与多义性，计

(下转第 196 页)

表 3　浏览路挖掘算法比较

	2-项浏览偏爱子路径挖掘算法	
	I	II
Pi=1	{AB,AC,AE,BC,CD,DF,EF,FB}	{AB,AE,BC,BF,CD,DE,EF,FB}
Pi=1.2	{AB,AE,BC,CD}	{AB,AE,BC,CD,DE}
Pi=2	{BC,CD}	{AE,BC,CD}
Pi=3	{/}	{/}

表 3 列举出浏览偏爱子路径挖掘算法在当初始条件相同(相关页面集都为{ABCDEF})时不同的浏览选择偏爱阈值下得到的不同浏览偏爱路径集。I 为文献[7]的实验结果,II 为本文的实验结果。通过表 3 可以看出,以本文中改进的浏览兴趣度为基本元素的算法 II 得到的偏爱路径比算法 I 准确。

由实验结果分析,本文算法在有效性和准确性上有一定的优势,可扩展性良好。

结束语　本文提出了一种改进的基于 Web 日志的用户浏览偏爱路径的挖掘方法。本文主要是在以单元数组的存储结构为基础建立的两个矩阵模型上,挖掘了不同的相似用户群体的相关页面集的浏览偏爱路径。方法只需访问一次数据库文件,减少了 I/O 负担。本文相对于其他算法,在有效性和准确性方面具有一定优越性,能准确、充分地表现不同用户群体在浏览路径上的偏爱倾向,可扩展性好。笔者在下一阶段将在浏览兴趣度及算法的处理对象上做进一步的研究。

参 考 文 献

[1] Hua Jiawei, Kamber M. Data Mining Concepts and Techniques [M]. China Machine Press, 2001

[2] Anand S S, Patrick A R, Hughes J G. A Data Ming Methodology. Cross Sales Knowledge Based System Journal, 1998, 10(7): 449-461

[3] Pierrako S D, Paliouras G. Web Usage Mining as a Tool for Personalization: A survey[J]. Kluwer Academic Publishers, 2003: 311-372

[4] XING Dong-shan, SHEN Jun-yi, SONG Qin-bao. Discovering Preferred Browsing Paths from Web Logs[J]. Chinese Journal of Computers, 2003, 11(26): 1518-1523

[5] NING Xiao-hong, YU Sen-sen. Study on s-Tree Algorithm for Personalized Recommendation [J]. Computer Science, 2007, 34(4): 217-221

[6] Zhang Hai - yu , Liu Xiao - xia . A New Way to Discover User Browsing Mode[J]. Computer Applications and Software, 2007, 24(2): 143-150

[7] Du Jia-qiang, Han Qi-rui, Wang Ke, et al. A Fast Algorithm for Mining User Frequent Paths from Web Logs[J]. Computer Engineering and Applications, 2005, 22: 164-167

[8] Tia Chang-peng. Base on the Analysising and Researching Web-sever Log of Web Qos[J]. Computer Science, 2007, 34(6): 78-80

[9] Mao Guo-jun, Duan Li-juan. Data Mining Principles and Algorithms[M]. Tsinghua University Press, 2005

(上接第 157 页)

算机只得采取"以数量换质量"的策略,即用多个明确的、单义的概念(及其关系)表达一个人脑中的概念(及其关系)。由此导致了概念系统规模的急剧增大,这不仅涉及工作量,更关键的是如此多的概念(及其关系)由谁来选取与甄别?只有领域专家,这在实际操作上有诸多困难。

(2)持续性维护。概念系统一定是开放的。人脑中的概念是通过后天习得和人际交互的传授(最典型的是教育),概念的更新与概念系统的丰富则是由其主动积极的思维活动来实现。相比之下,当今的计算机只有"人操作的外部输入"这一种途径,因此必须不断地由人对概念(及其关系)进行添加、修改与更新。

(3)分布式及其集成应用。人类知识的巨量与进化决定了概念系统只能是分布式的。多个概念系统存在的优势是兼顾不同领域、并行开发与维护、使用时的按需取舍等。然而,分布式带来的直接问题是如何相互调用,其中异质性(heterogeneity)是最大阻碍。本体集成(ontology integrating)应运而生,包括本体映射(ontology mapping)、本体合并(ontology merging)、本体调整(ontology aligning)和本体连接(ontology articulating),成为目前 Ontology 研究中的主要分支。

结束语　语义 Web 所说的"语义"就是组成计算机中的概念系统的那些概念。如此界定只是一种现实可行的选择,所做的两项简化为"搁置语境"和"以概念近似观念"。

语义 Web 由 Web 扩展而来,两者同为信息集合。前者中的信息采用概念标记符标记,供机器交互理解,后者中的信息仅用格式标记符标记,由人直接阅读。扩展的目的就是要从"由人直接阅读信息内容"进步到"先由机器阅读并进行推理,再将结果按需提供给人使用"。扩展的措施为(1)信息采用概念标记;(2)计算机内置概念系统。

概念标记就是用概念标记符(即概念系统中的概念)对将要交由计算机处理的信息进行标记。也就是用 XML 对信息做"语义化处理"。目前的难点主要有:信息内容的切分、概念标记符的选用、标记过程的自动化等。

概念系统就是 Ontology,是由概念、概念的属性、概念之间相互关系组成的层次网络。它的主要功用有二:读出信息的直接语义、进行概念推理。基于概念系统通达性的概念推理才是语义 Web 的最突出的特征。应用概念推理计算机能够读出给定信息的多种语义。构建概念系统目前需要解决的问题主要有:规模庞大、持续性维护、分布式及其集成应用等。

参 考 文 献

[1] Berners-Lee T, Hendler J, Lassila O. The semantic Web[J]. Scientific American, 2001, 284(5): 34-43

[2] Jacobs I, Walsh N. Architecture of the World Wide Web[R]. W3C: Recommendation. http://www.w3.org/ TR/2004/REC-Webarch-20041215/

[3] Uschold M. Where are the semantics in the semantic Web[EB/OL]. http://www.starlab.vub.ac.be/WhereAre Semantics-AI-Mag-FinalSubmittedVersion2.pdf

[4] Antoniou G, Harmelen F. The semantic Web primer[M]. London: The MIT Press, 2004: 17-18

[5] 张志毅,张庆云. 词汇语义学(修订本)[M]. 北京:商务印书馆, 2005: 1-3

[6] 李幼蒸. 理论符号学导论[M]. 北京:社会科学文献出版社, 1999: 128-133, 292-296

[7] 陈嘉映. 语言哲学[M]. 北京:北京大学出版社, 2003: 44-57

[8] Frege G. 弗雷格哲学论著选辑[M]. 王路,译. 北京:商务印书馆, 2006: 95-119

[9] 王寅. 语义理论与语言教学[M]. 上海:上海外语教育出版社, 2001: 34-38

[10] 罗旋. 基于复句领域本体的语义标注方法研究[D]. 硕士论文. 武汉:华中师范大学, 2006

[11] Gruber T R. A translation approach to portable ontology specifications, KSL92-71[R]. San Francisco: Knowledge Systems Laboratory of Stanford University, 1993