

信息检索

HTML+⁺⁺

WWW Internet

(5)

计算机科学2000 Vol. 27 No. 2

57-60

HTML⁺⁺——一种增强 WWW 智能的途径HTML⁺⁺—An Approach to Promote the Web with Intelligence

李立新 张为群

(西南师范大学计算机科学系 重庆400715)

G354.4

TP393

Abstract There are problems to access information from the Web. It is the HTML's weakness in knowledge representation and the WWW's shortage in knowledge handling. This paper presents a object-oriented hyper text makeup language—HTML⁺⁺, which is an enhanced version of HTML, in order to optimize the information structure of hyper text and provide it the capability of semantic representation. HTML⁺⁺ provides WWW a practicable approach of importing the capability of information processing such as reasoning. This makes it possible accessing information in an ad-hoc, direct, precision and customized style.

Keywords Information processing, Computer network, Object-oriented, WWW, HTML⁺⁺

1 引言

Internet,特别是万维网 WWW 的迅猛发展,为访问巨大数量的信息,尤其是以超文本标记语言(Hyper Text Make-up Language,简称 HTML)书写的文档成为可能。WWW 在信息定位和定界上存在着单步定位和偏差定界的问题,这既造成了信息获取困难,也造成了冗余信息传输的浪费。

目前这一问题已成为关注的焦点之一。有许多工作(例如文[5])通过引入智能机制来提高信息检索效率,但所关心的对象是对象数据库这样的良结构信息。也有工作(例如文[4])通过扩展 WWW 来减少文档传输开销,但它是将文档为基本单位的。目前,Internet 上的信息正越来越多,且多是以 HTML 书写的,因此研究 HTML 文档中信息的高效获取十分重要。本文认为,造成在 WWW 上获取信息困难的主要原因是 WWW 不“理解”它所处理的信息,而且也无法对其内容作出自动处理,这无疑增加了客户端用户手工操作的负担。HTML 书写的文档中有大量的知识,但这些知识是隐含在图形、文字等媒体中的,必须将这些知识有效地表示出来,才能做到信息获取的自动化处理。本文分析了 WWW 的工作机制,介绍了超文本描述语言 HTML⁺⁺及其在信息自动获取和组织中的作用。

2 WWW 上的信息获取

WWW 是受 TCP/IP 支持的应用协议集,采用 URL、HTTP、HTM 和 CGI 等协议和标准进行信息的

定位、存取和显示。WWW 已成为 Internet 上最主要的应用之一,有无数的以 HTML 书写的信息传输于世界各地之间。

2.1 WWW 的信息结构

WWW 可被静态地看作是一个全球范围的相互链接的多媒体信息网。也可被逻辑地视为一个用 HTML 书写的由分布在许多物理站点上的超文本文档通过热键互连而成的巨大的分布式超文本结构。作为 Internet 资源的文档都有一个被称为 URL 的 Internet 地址,热键实际上就是有一定 URL 作为其值的 Internet 地址指针。一个文档中的任何一处都可以有一个指向另一文档或者其自身某处的热键。URL 协议定义的 URL 的实例如:<http://www.cs.swnu.edu./tutor/ct.html>

2.2 WWW 的工作机制

WWW 采用了“客户机/服务器”计算模式,其间依照 HTTP 协议交换信息。用户就是通过客户机(浏览器),借助这种信息交换来获取信息。客户机要实现以下功能:

- 1)与服务器建立连接,并用 URL 向服务器索取文档。
- 2)按要求的方式排版显示信息。
- 3)搜集用户输入,做出响应,并在必要的时候将其交服务器进行进一步处理。

服务器是具有以下能力的一个运行中的软件:

- 1)侦听来自用户端的连接请求。
- 2)依据 HTTP 协议与客户机会话以及传送信息。

3)依据客户端的要求进行必要的信息处理工作。

应当指出的是客户机能对用户作出的反应是有限的,同时也难于扩展新的相应。服务器能作出的信息处理也是很有限的。CGI 协议规定了这种交互方式^[1]。

2.3 WWW 上信息获取的过程与方式

WWW 上的信息获取是在导航机制的协助下,依照信息地址实现的。显然,在已知信息地址的情况下这种方式是快捷的。另外,文件级的文档是 WWW 获取信息的基本单位。这些特点在信息定位和定界上引发了以下问题:

1)单步信息定位:导航只允许沿着热键一步一步地追踪信息。在这个追踪过程中客户机接收了许多最终被排除的信息。

2)非即席信息查询:只有有限的服务器提供信息查询服务或代理这种服务。查询需要用户在服务器间来回切换。这也是造成迷失和冗余信息传输的一个原因。

3)偏差信息定位:客户机在得到了一些不需要信息的同时只得到所需信息的一部分,究其原因 WWW 既不能排除一个文档中对用户无用的部分,也不能同时从不同的文档中获取信息,即,WWW 不能定制信息。由于文档是由提供者按自身设想制作的,因而这种偏差往往很严重。

3 HTML++ 超超文本标记语言

超超文本是一种在逻辑结构上受“语义网络(OSN)”规范的具有复杂热键联系的文本。超文本是超超文本的一种特例。HTML++ 是一个超超文本标记语言,是 HTML 的超集,其上扩充了类定义、对象标识、热键标注等。

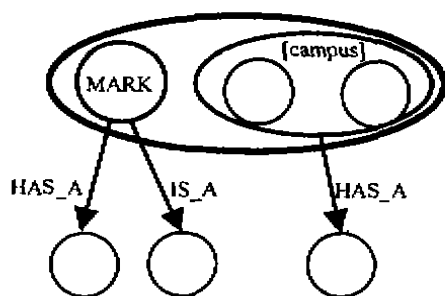


图1 对象语义网络

3.1 语义网络 OSN 表示方法

在 OSN 表示方法中,一个 OSN 是由一些以有向图表示的三元组 (Object1, Arc, Object2) 连接而成的 (图1)。其中对象表示事物,弧表示对象间的关系。例如

三元组 (University, HAS-A, Computer-Dpt.) 表示 University 大学有一个 Computer 系。弧的标注可以有 IS-A, HREF, PART-OF 等多种。

OSN 是超超文本的抽象。例如图1所示的 OSN 即是对图2所示超超文本的抽象。OSN 与超超文本的术语是相关的,表1列出了 OSN 与超超文本不同术语间的对应关系。

表1 OSN 与超超文本的比较

OSN	超超文本
对象	元素
对象类	类
弧	外部属性/热键
对象标识	元素标识/EURL
环境对象	环境元素
复合对象	复合元素

3.2 类的定义

图2所示超超文本的第三行到第十行给出了一个 IMGX 类的定义。IMGX 被定义为 IMG 的子类,它从 IMG 中继承了一些属性和方法。第四、五行定义了两个不同类型的属性,属性 ACTION 可以在该类实例化时接收初始化值,从而实现实例的初始化。缺省的属性类型是私有的,第6行至第9行定义了一个名为 OnLBt-nClk 的方法,相应“LBt-nClk”为消息。第7、8两行定义了该方法的行为,其中单引号引起的均为该类或其超类的属性。

```

1 (HTML)
2 (HEAD)HTML++ Illustration(/HEAD)
3 (CLASS NAME="IMGX" APPEND="IMG")
4 (ATTRIB NAME="ACTION" TYPE="INTERFACE")
5 (ATTRIB NAME="POS")
6 (METHOD NAME="OnLBt-nClk" MSG="LBt-nClk")
7 (LOCALEXEC NAME="GetPos" RETURN="POS")
8 (REMOTEEEXEC NAME="ACTION" RETURN="SRC" PARAM="SRC,POS")
9 (/METHOD)
10 (/CLASS)
11 (BODY)
12 # MARK (IMGX SRC="MARK.GIF" ACTION="IMG/CGI.EXE"
13 ISA="http://www.cs.edu/a.htm#flag"
14 HASA="http://www.cs.edu/a.htm/[map]#[star]")
15 # [campus] (IMGX SRC="map.gif" ACTION="cgi.exe" HASA="[dpt]")
16 (B)SWNU Campus Map(/B)
17 Click the map for details
18 (/IMGX)
19 (/BODY)
20 (/HTML)
    
```

图2 超超文本文档实例

3.3 元素及其标识

在 HTML++ 中,类的实例被称为元素。图2中第12行到第14行以及第15行到第18行分别定义了一个简单

元素(图1中的 MARK 元素)和一个复合元素(图1中的 [campus] 元素)。第16行和第17行上的是该复合元素 [campus] 的两个寄居元素,外层上的元素为这两个寄居元素提供了环境而被称为它们的环境元素。

在第12行到第15行两元素的开始之前都有用于标识它们的标识 # MARK 和 # [campus]。其中 # MARK 是一个标号标识,它表明了一定的地址,用来实现“静态连接”。而 # [campus] 是一个关键字标识,是抽取了元素知识特征的符号,它代表了一类对象元素的共同属性,通过它可以实现“动态连接”。一个元素可以有多个关键字标识,用以刻画它不同方面的分类特征和语义特征。

3.4 热键

图2中第13行和第14行上的 ISA 和 HASA 都是从超类继承下来的特殊属性,即外部属性。在 HTML⁺⁺ 中也称其为热键。ISA 和 HASA 及与 HTML 相同的 HREF 在 HTML⁺⁺ 中有不同的语义内涵,表达了多种语义关系。

3.5 扩展 URL(EURL)

EURL (Extend URL) 通过以下扩展支持超超文本结构:

1) 为以文档为环境的元素定义 URL。这使得所有元素都能被任何其它元素引用。

2) 将唯一的元素地址扩充为唯一的元素集合地址,这种地址被称为查找地址,而 URL 地址被相应地称为静态地址,例如,图2中第13行上的 EURL:

"http://www.cs.edu/a.htm#flag"

是一个引向某个文档内部某一元素的静态地址。而14行上的 EURL: "http://www.cs.edu/a.htm/[imap]#[star]" 是一个引向一些文档内部的一些元素的查找地址,它表示的是一个对象与一类对象之间的知识关联。客户点击该 EURL 后会引起服务机制的推理和检索过程(见4节)。

表2 HTML⁺⁺ 与 HTML 的比较

HTML	HTML ⁺⁺ 兼容	HTML ⁺⁺ 扩充
部分元素可嵌入	全部元素可嵌入	
部分元素有热键	全部元素有热键	
元素	无标识元素	元素
标号		元素标识
热键	静态热键	动态热键
HREF	HREF 式热键	各种语义标注
		自定义类

3.6 HTML⁺⁺ 与 HTML

HTML⁺⁺ 在完整地保留了 HTML 语法和语义的基础上进行了扩充,从图2中可以看到 HTML⁺⁺ 对

HTML 语法的兼容性及扩充语法风格与 HTML 语法风格的一致性。表2描述了这两者在语义上的联系。

综上所述,HTML⁺⁺ 是一个具有描述 OSN 能力的 HTML 超集。用 HTML⁺⁺ 描写的超超文本能有效地表示知识以便能为机器处理。

4 HTML⁺⁺ 与信息的自动获取与组织

HTML⁺⁺ 文档中可以包含显式的知识,这种信息组织形式有利于信息的自动获取与组织。在推理机制的作用下,利用这些显式的知识可以导出新的信息。这个推理机制实际上是以超超文本为基础的分布式环境下 OSN 推理机制——InfoServer。与语义网络推理机制类似,InfoServer 主要依靠继承和匹配推理。但是 InfoServer 的主要目标是从已知信息 A 按照它与位置信息 X 的关系 r 导出 X。

InfoServer 采用增量式逐步求精的求解策略来获取未知信息 X,即逐步由粗到精、由少到多地完善解答。它从当前已知信息 A 出发,逐步向深度(包括元素的嵌套深度和热键连接深度)推进。每一步推理都引起搜索历史的变迁和当前目标的变化,并且从当前元素开始依据当前历史记录和目标寻求一些能使目标范围缩小的语义连接或嵌套指示的路径以继续搜索,直到目标被尽可能多地满足。例如,我们可以从包含有中国各大学的网页上要求查找各大学的计算机系的有关信息。InfoServer 在接到这个请求后会利用网页中的关系 (University, HAS-A, Computer-Dpt.) 以及 InfoServer 规则库中的规则 (University, HAS-A, X-College)&(X-College, HAS-A, Computer-Dpt.) = (University, HAS-A, Computer-Dpt.) 等对各大学甚至其下属学院进行搜索。

在搜索过程中可能会有多个相同元素出现(例如有多个元素满足查找连接),这时搜索过程会按照知识的结构建立归并框架,将它们归为一个对象,然后再对它展开搜索。例如可能的搜索中间结果是:

Departments of Computer in China (from CERNET)
University: Tsinghua University
Information school
Dep. Of Computer Science and Technical
(is coming...)
University: Beijing University
Dep. of Computer Science
...
University: Nankai University
(is coming...)
University: (is coming...)

服务器在得到来自客户的求解目标后,不但要对它进行分析,部分求解和分解,与此同时,它还形成解答框架交给客户机,并将子目标及其解答在上层框架中的位置一起交给下一轮求解。这些子目标及其解答在上层框架中的位置被返回到客户机,以便客户机将

UML 面向对象 Object-Z 对象模型

16

60-63 基于 Object-Z 的 UML 对象模型的形式化

The Formalization of Object Model in UML Based on Object-Z

杨卫东 蔡希尧

TP312UM

(西安电子科技大学软件工程研究所 西安710071)

Abstract UML is the main visual Object-oriented modeling language currently, which is used widely and supported by most CASE tools. Comparing with traditional Object-oriented methods, UML describes its semantics and syntax more rigouly by using metamodel and Object Constraim Language. But some important concepts in UML are not specified clearly. This paper presents a formal specification for object model of UML, mainly includes the concepts of class, association, association class, aggregation, and inheritance, etc, so that the analyse, verification, refine, and consistent cheking can be applied to object model.

Keywords Object-oriented, Formal specification, UML, Object-Z, Object model, Schema

1 引言

UML 作为面向对象的可视化建模语言,已被对象管理集团(OMG)作为面向对象分析和设计的标准,获得了众多工具的支持。UML 提供了不同抽象层次的描述以支持面向对象的分析、设计和实施,它从不同的视图描述软件系统,减少了建模的复杂度,更为重要的是建立了基于元模型的体系结构,提供了较为灵活的扩充机制,使开发人员可以根据不同的领域需求定制 UML,也易于加入新的建模概念。

UML 与以前的主流软件开发方法(OMT、BOOCH 方法和 OOSE)相比,其中一个显著特点就是加强了建模语言语义的严格描述,主要使用了元模型、对象约束语言(OCL)和自然语言。元模型的使用严格地描述了 UML 建模技术的抽象语法,可用于语义描述的基础。OCL 适合于表达约束规则,但用于描述动态语义则不容易被理解,也缺乏语义的形式描述。因此,UML 虽然在建模概念的语法和语义的描述上进

了一步,还是难以支持软件模型的分析、验证以及一致性检查等。

Object-Z 是基于 Z 语言的面向对象的扩充,与 OCL 不同,Object-Z 和 Z 具有严格的语义,提供了较强的证明、分析和验证的能力,具有众多支持工具。Object-Z 支持主要的面向对象的概念(类、继承、多态等),用 Object-Z 作为 UML 模型的形式化描述语言,可以在软件开发的初期阶段分析模型、检查错误,提供了验证模型和开发过程中严格求精的基础,能够支持模型的一致性检查等。

对象模型是 UML 中最重要的模型,本文用 Object-Z 描述了 UML 对象模型的主要概念,如:类、关联、聚合、继承等。

2 UML 对象模型的形式化

UML 从四个视图描述软件系统,即使用实例(use case view)、逻辑(logical view)、物理(physical)、交付(deployment view),逻辑视图主要描述了软件的逻辑

它们逐步整理到当前解答框架中。

参考文献

- 1 Handley M, Crowcroft J. The World Wide Web. UCL Press,1995
- 2 Duntman J. Mosaic Pocket Companion. Coriolis Group Inc., 1995
- 3 Kuokka D, Harada L. Matchmaking for Information A-

gents. IJCAI-95

- 4 Bestavros A. Demand-based document dissemination to reduce traffic and balance load in distributed information system. In: Proc. of the 1995 7th IEEE Symposium on Parallel and Distributed Processing. IEEE,1995
- 5 Newton P. Capitalizing on the World Wide Web's Knowledge Infrastructure. 12th Conf. Proc. ADBS-IDT,1995