

16-18

对象语义网络: WWW 信息结构模型

OSN: A WWW Information Structure Model

李立新

(西南师范大学计算机科学系 重庆 400715)

Abstract There is more and more information available in the Web. There are also problems we encountered to acquire the information. One of the reason is that the Web lacks knowledge representing capability. This paper presents a WWW information structure model called "Objective Semantic Network(OSN)", with knowledge representing capability, and a corresponding reasoning mechanism for the Web. This makes it possible to acquire ad-hoc, direct, and precision location and customization information.

Keywords Information process, Computer network, AI, Object oriented, Internet, WWW, Knowledge representing, OSN

1 引言

基于 HTTP 的 WWW 在信息定位和定界(information locating and narrowing)上存在着单步定位和偏差定界的问题。这些问题在造成信息获取困难的同时,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[4])通过引入智能机制来提高信息检索效率,但是它关心的对象为数据库这样的良结构信息。也有工作(例如文[3])通过扩展 WWW 来减少文档传输开销,但它是将文档作为基本单位的。由于 Internet 上已经存在有大量的和不断增加的以 HTML 书写的信息,因而研究 HTML 文档中信息的高效获取是有意义的。

本文认为造成 WWW 上获取信息困难的主要原因是,WWW 既不“理解”它所处理的信息,也无法对其内容作出自动处理,从而增加了客户端用户手工操作的负担。HTML 写书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。

本文分析了在 WWW 上获取信息的过程与方式中存在的问题,针对它建立了一种称为“对象语义网络(OSN)”的知识表示方法。并介绍了 OSN 相应的推理机制,即 InfoServer,它是对 WWW 的扩充。

2 WWW 上的信息获取

WWW 是受 TCP/IP 支持的应用协议集。它采用 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 书写的巨大的分布式超文本结构。该超文本结构是由分布在许多物理站点上的超文本文档通过热键(Internet 地址指针)互连而成的。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。

2.2 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的一部分信息。其原因在于 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 目前还不能定制信息。由于文档是由提供者按自身设想制作的,并没有一个规范,所以这种偏差往往不能控制,而且很严重。

3 对象语义网络(OSN)

对象语义网络(Objective Semantic Network,简称 OSN)表示法是由语义网络表示法和面向对象表示法结合而成的一种混合知识表示方法。在后面的讨论中可以看到,OSN 表示法能适应为 WWW 增强智能的需要。

3.1 OSN 结构

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,例如图 1 中的(A, IS-A, F)。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 HAS-A, IS-A, PART-OF, HREF 等多种。

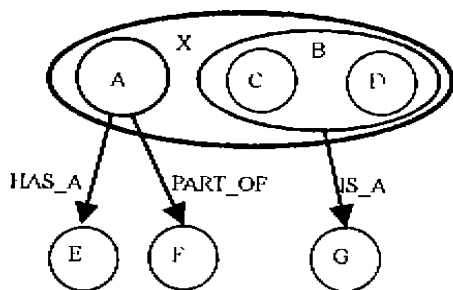


图 1 对象语义网络

3.2 OSN 对象和 OSN 对象类

OSN 对象由一些属性和方法封装而成。与面向对象的方法相似,方法分为公有和私有两种。同样地,具有相同属性和方法的对象属同一个对象类。类间可以继承进而形成类层次。系统提供 IMG、B、HREF 等一些基本对象类,并允许从这些基本对象类派生自定义对象类。例如我们可以从 IMG 类派生 IMGX 类。方法可以用来表述对象类的行为,例如我们可以为一个图形对象类 IMGX 定义 OnLBtnClick 方法以响应单击鼠标左键事件。其中 HREF 及其派生类的某些方法可以用来刻画信息检索行为以及从已知信息获得未知信息的推理行为,这使得信息的表示与信息的获取以及组织能有机地结合起来。

3.3 外部属性和内部属性

OSN 对象属性有内部属性和外部属性两种。内部

属性相当于面向对象表示方法中的属性,用于表示对象自身的性质,如定义图形对象的大小等。外部属性则用于描述对象间的语义联系,如 IS-A, HREF, PART-OF 等。外部属性只能由相应的推理机制访问。OSN 中所有对象共享相同的外部属性类。

3.4 环境对象与寄居对象

OSN 方法不支持嵌套的类定义,但是可以通过对象“寄居”允许更为灵活的对象嵌套。对象甲寄居在对象乙中是指对象甲自动获得了对象乙中的那些对象甲所下具备的属性和方法,同时对象甲的那些与对象乙同名的属性和方法将被对象乙的属性和方法取代。这种方式与生物界中生物受环境影响的情况类似。对象甲被称为对象乙的一个“寄居对象”,对象乙被称为对象甲的“环境对象”。例如图 1 中的对象 C 和 D 是对象 B 的“寄居对象”,而对象 B 是 C 和 D 的“环境对象”。很明显,多层次的寄居将使环境与寄居成为一对相对的概念。例如图 1 中的对象 B 同时也是对象 X 的一个“寄居对象”。寄居对象与它所寄居的环境对象也可以被看作是一个整体,并被称为“复合对象”。图 1 中的对象 B、C 和 D 的整体构成了一个“复合对象”,而那些不是任何对象“环境对象”的对象则被称为“简单对象”。比如,图 1 中的对象 A、C、E 等都是“简单对象”。

3.5 OSN 表示法及其它表示法

语义网络法与 OSN 法均用弧表示关系,用结点或对象表示事物。然而 OSN 对象可以有更复杂的嵌套结构和丰富的方法。它是一个多层次、多深度的三维复杂网络。

与面向对象表示法相比,OSN 对象的外部属性允许表达更丰富的对象间语义联系,同时对象寄居也使 OSN 有更好的开放性。

4 信息的自动获取与组织

本节以超文本为基础讨论分布式环境下 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。

4.1 “客户机/服务器组”计算模型

InfoServer 运行在分布式环境下,它采用了成为“客户机/服务器组”的这种由“客户机/服务器”推广而来的计算模型。亦即在客户机提出一个求解要求后,将会有一组服务器协调求解,联合作答。构成服务器组的服务器不是固定不变的。

4.2 增量式逐步求精的求解策略

未知信息除了其细节是未知的以外,其数量和结构也是未知的。InfoServer 采用增量式逐步求精的求

解策略:逐步由粗到精地在深度上和由少到多地在广度上完善解答(图2)。应注意的是深度和广度是多个层次的相对概念,因而这种策略有好的并发性,适合“客户机/服务器组”计算模型。

搜索中间结果	注
Departments of Computer in China	X
University:Thinghua University	A
Information school	E
Dep. Of Computer Science and Tec.	a
(is coming...)	
University:Beijing University	B
Dep. of Computer Science	b
...	
University:Nankai University	C
(is coming...)	
University:(is coming...)	

图2 有组织地获取中的信息

4.3 深度搜索和广度归并

InfoServer 为获取未知信息 X,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继续搜索,直到目标被尽可能多地满足。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。图2是在客户要求查找各大学的计算机系的有关信息后 InfoServer 进行搜索所得到的中间结果。图3展示了图2对应的搜索状态。这些搜索过程是在元素所在服务器上独立进行的。

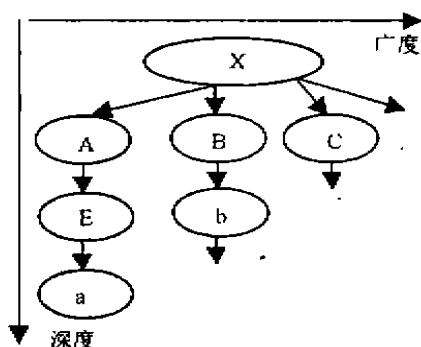


图3 搜索状态

4.4 推理控制与解答组织

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一边进行推理一边搜索。在 InfoServer 中,推理控制与解答组织是相关的。服务器在得到来自客户的求解目标后,对它进行分析,部分求解和分解,与此同时,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,以便客户机将他们逐步整理到当前解答框架中(图2)。

4.5 InfoServer 的特点

与 WWW 相比,InfoServer 有以下特点:

- 1)通过查找热键的支持,实现在现行信息环境中的即席查询。
- 2)通过继承,使客户机不经过中间过程即能得到目标信息,实现直接定位。
- 3)以元素为查询单位可以最大限度地排除冗余信息;通过目标被尽可能多地满足搜索,又使客户端能得到完整的信息。这两种处理使 InfoServer 实现了精确定位。
- 4)通过在深度上的信息缩写和广度上的综合,提供定制信息。

5 OSN 的描述语言——HTML⁺⁺

超超文本是一种在逻辑结构上受 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。HTML⁺⁺是在 HTML 上扩充类定义、对象标识、热键标注而形成的。

参考文献

- 1 Duntelman J. Mosaic Pocket Companion. Coriolis Group Inc., 1995
- 2 Kuokka D, Harada L. Matchmaking for Information Agents. IJCAI-95
- 3 Bestavros A. Demand-based document dissemination to reduce traffic and balance load in distributed information systems. In: Proc. of the 1995 7th IEEE Symposium on Parallel and Distributed Processing. IEEE, 1995
- 4 Newton P. Capitalizing on the World Wide Web's knowledge infrastructure. In: 12th Conf. Proc. ADBS-IDT, 1995