

基于扩展标记图的虚拟网页技术^{*}

The Designing Technique with Virtual WWW Page Based on Extended Tag Graph

朱征宇 朱庆生 王 茜

(重庆大学计算机学院 重庆400044)

Abstract By introducing the concepts of original material database and extended tag graph, this paper presents a new idea--any HTML document can be designed indirectly by a virtual WWW page. With the traditional method, a WWW page is often given out by an HTML document (called an actual page) which contains both the structure and the contents of the page. But with the designing technique of virtual WWW page, the contents of the page and the structure of the page have been separated. The HTML document of a page will be given out indirectly with a Virtual WWW page which contains only the structure of the page without actual contents in it. The actual contents of all the pages are laid in an original material database. This new technique plays an important role in research on reducing redundant data among HTML documents, uniformity maintenance of the contents of different HTML documents, block-based design for WWW pages and the query language of searching for Web information.

Keywords HTML document, Tag graph, Original contents database, Virtual WWW page

大量的 Web 信息资源采用 HTML 文档形式存放,带来诸多不利因素。其一,HTML 文档因包含具体数据内容,不可避免地会涉及到数据冗余和一致性维护的问题。比如,在一个电子商场网站设计中,为服务于不同客户群,一双皮鞋的基础数据(厂家介绍、产品说明和图片等)可能会重复地放入多个网页中,如按分类、按品牌和按厂家等方式组织的网页。其二,HTML 文档不利于构建新网页的重用,难以适应变化需求,如当有新客户群希望按老中幼年年龄段来浏览商品时,新网页须从头组织,且可能产生新的数据冗余。其三,HTML 文档的信息检索和查询优化较为困难。

然而,关系数据模型中由于采用基表和视图技术,使类似问题得到很好解决。由此得到启发,提出了虚拟网页设计技术。通过建立素材库和定义扩展标记图,将网页的数据组织结构和具体内容分离,使问题得到较好解决。虚拟网页技术,还有利于借鉴标记图^[1]、OEM^[2]和 OIM^[3]等数据模型已有研究成果,有利于 Web 查询语言和信息检索优化的研究。

1 标记图简介^[1]

文[1]指出,HTML 文档中存在大量标记,可以利

用这些标记所携带的语义信息来构建 HTML 文档的标记图模型。

首先,通过引入唯一标识来区分 HTML 文档中的标记,并通过定义 HTML 文档中标记间的位置先后关系和包含关系、标记的一步可达标记集等基本概念,给出了 HTML 文档的标记树概念。然后,通过考虑文档内和文档间的链接(通过标记 A 和标记 LINK 实现的),分别引出了 HTML 文档的标记图和标记超图概念。

标记树实际上是将标记的包含关系和标记附带的属性转化为树结构中的父子关系构成的,父节点(标记 t)有若干条指向子节点(t 的一步可达标记集,即标记 t 直接包含的所有下一层标记)的有序带标识边,边的顺序用于反应标记在文档中的顺序关系。此外,父节点(标记 t)的若干属性值也分别作为子节点,顺序号都是 0,因为属性不考虑次序。特别地,把标记 t 的内容(即标记 t 的首尾标记之间所包含的字符串)当作标记 t 的一个特殊属性 Content 看待,也作为一个子节点。进而引入文档内链接将标记树推广为标记图。

为反映文档间的关系,可将文档间的超链接引入标记图,构成多文档的标记超图。文档 $d, d_1, \dots, d_n$ 的

^{*} 本文研究工作得到国家863计划项目支持(863-511-910-101-2)。朱征宇 副教授,主要研究领域:电子商务、WWW 智能检索和数据库技术。朱庆生 教授,主要研究方向为多媒体数据传输与压缩技术。王 茜 副教授,主要研究领域:商务智能及电子商务安全技术。

标记超图,是 $d_1, d_1, \dots, d_n$ 各自标记图的并集,在通过文档间的超链接连接而成。

标记图是一个带根连通边有序有向图,精确地反映了 HTML 文档的内部结构,而且很容易从标记图逆向生成 HTML 文档,即 HTML 文档与其标记图之间是相互唯一确定的。

2 扩展标记图

注意到 HTML 文档与其标记图之间是相互唯一确定的,我们可以换个角度来考虑问题,在设计一个网页时不是给出其 HTML 文档而是它的标记图,但仅仅如此还不够,因为标记图中不仅包含了数据组织结构,还包含数据,它与相应的 HTML 文档仅仅是表现形式不同。

为了实现真正意义上的虚拟网页设计技术,需要对标记图进行必要的改造,引入扩充标记图。对标记图的改造原则有三:一是移去数据内容仅保留结构。二是减少不必要的标记节点,简化标记图。三是有利于网页设计的模块化和重用。主要改造内容如下:

(1) 标记图的节点,扩充为一个五元组(标识,名称,类型,属性,值地址),分别表示该节点对应标记的唯一标识、标记名称、标记类型、标记的属性和标记内容 Content^[1](该标记的首尾标记之间包含的字符串)的地址。因采用值地址来定位内容位置(在后而称为素材库中),使得整个标记图中不再存放网页的具体内容,而仅描述其信息组织结构,由值地址定位的内容,一般是标记的内容 Content,也可能是一个 URL 链接(到另一 HTML 文档、另一标记图等)。

(2) 标记的属性都不单独作为子节点。一般属性包含在标记节点的属性内,但表示文档链接的属性(如标记 A 的 href 属性、img 的 src 属性、Applet 的 code 属性等)将作为节点的内容包含在值地址内。

(3) 支持扩展标记。通过引入扩展标记,允许把一个 HTML 文档描述段,即用 HTML 语言编写的一段 HTML 文档内容,但不一定是一个完整的 HTML 文档(以下简称 HTML 标记段),看作标记图的一个节点,而把该标记段当作扩展标记的内容来看待。例如,可把 <from> 标记段、<table> 标记段、或下述 HTML 标记段都可看作一个节点。

```
<select name="s2" multiple size=3>
  <option>1
  <option>2
  <option>3
</select>
<ol type="I">
  <li>one o'clock
  <li>two o'clock
</ol>
```

(4) 支持标记图的嫁接。允许把一个标记图(子图)

当作另一标记图(母图)的节点,当然应符合 HTML 语言规则,即子图与母图嫁接后仍反映一有效 HTML 文档的结构。

(5) 支持到标记图的热链接。对属性 href 节点,允许接受标记图的 URL 地址,类同于 HTML 文档中的 URL 热链接。

经过上述改造后的带根连通有向图,本文称为扩展标记图。在上述分析的基础上,下面给出扩展标记图的严格定义。

设有向图 $G(V, E)$, V 是 G 的顶点集, $E = \{ \langle v_i, v_j \rangle \mid v_i, v_j \in V \}$ 为 G 的边集,这里 $\langle v_i, v_j \rangle$ 表示从顶点 v_i 到顶点 v_j 的有向边。

定义1 设有向图 $G(V, E)$, $v, w \in V$, 若存在 $k_1, k_2, \dots, k_i \in V$ 使 $\langle v, k_1 \rangle, \langle k_1, k_2 \rangle, \dots, \langle k_{i-1}, k_i \rangle, \langle k_i, w \rangle \in E$, 则称 v 到 w 是可达的, 而称 $P_{vw} = \langle v, k_1 \rangle \langle k_1, k_2 \rangle \dots \langle k_{i-1}, k_i \rangle \langle k_i, w \rangle$ 为 v 到 w 的一条有向路径。

定义2 若有向图 $G(V, E)$ 中存在一顶点 $v \in V$, 且 v 到任何 $w \in V - \{v\}$ 都是可达的, 则称 v 为图 G 的完全可达顶点。

定义3 设有向图 $G(V, E)$, 若 G 中存在且仅存在一个完全可达顶点 r , 则称图 G 为带根可达连通有向图, 而 r 称为图 G 的根, 记为 $G_r(V, E)$, 有时简记为 G_r 。

定义4 在 $G_r(V, E)$ 中, 若顶点 $v \in V$ 不存在有向边 $\langle v, k \rangle \in E$, 则称 v 为 G_r 的叶节点, 否则称为 G_r 的非叶节点。用 V_0 表示 G_r 的叶节点集, 用 V_1 表示 G_r 的非叶节点集。

定义5 在 $G_r(V, E)$ 中, 对其任何顶点 v 的所有以 v 为起点的有向边, 都能通过序号进行区别, 则称 G_r 为边有序 G_r , 记为 $G_r(V, E^d)$ 。这是可记顶点 v 的第 i 条边 $\langle v, w_i \rangle$ 为 $\langle v, w_i, i \rangle$ 。

定义6 扩展标记图是用于描述 HTML 标记段的带根连通边有序有向图, 且是一个特殊的 $G_r(V, E^d)$, 其中 V 是由 HTML 标记符或扩充标识符为节点的集合, 每个节点都是一个五元组 (oid, n, t, a, c) , oid 为标识, n 为标记名, t 为标记类型, a 为标识的属性(组), c 为标记内容的值地址。边的序号表示标记在 HTML 标记段中的顺序关系。

显然对于扩展标记图与 HTML 标记段有: 一个 HTML 标记段一定有相应的扩展标记图(不唯一), 而一个扩展标记图唯一确定一个 HTML 标记段。

3 虚拟网页技术

3.1 技术思路简述

根据 OEM 模型^[3-4]、OIM 模型^[5]和标记图模型^[1]的研究可以看出, HTML 文档用半结构化数据模型来

描述是较为合理的选择,对信息检索、查询优化和 Web 上查询语言等问题研究起着重要作用,但对于 HTML 文档引起的信息冗余、一致性维护、适应不断变化的需要以及网页的快速设计等问题,并未得到解决。本文通过对这些半结构模型的分析,从关系数据模型得到启发,提出了虚拟网页技术的解决办法。

虚拟网页的主要技术思想是:将网页设计所需的所有基本信息放入一个素材库中,在设计一个网页时,一改传统的思维观念,不是采用 HTML 语言去直接编写一个实际的 HTML 文档,而是基于素材库上定义一个虚拟网页,来间接表示 HTML 文档。

一个虚拟网页,仅描述网页的信息组织结构但并不包含具体内容,内容存放在素材库中,但通过一个自动转换程序可以形成一个唯一确定的 HTML 文档。一般地,这样的 HTML 文档并不实际保存,只是在 Web 服务器响应客户浏览器发出对虚拟网页调用请求时才动态形成 HTML 文档,并返回给浏览器。这正是虚拟网页的真正含义。下面,对虚拟网页技术逐步加以介绍。

3.2 素材库

就象关系数据库中基表是创建视图的基础一样,在虚拟网页技术中,网页设计的基础是素材库,它包括一个网站制作所需的所有基础素材,诸如文字说明、HTML 标记段、图像文件和 VEScript 程序等。素材库可用关系数据库和文件目录来管理,也可仅采用文件目录来组织。其组织方式将在另文中讨论。

特别说明,素材库中也允许存放基于素材库定义的中间件。即素材库可以不断扩充,以便于网页的模块化制作、快速生成和应用维护。

3.3 虚拟网页设计

虚拟网页设计,实际上就是基于素材库的扩展标记图定义,扩展标记图采用何种文件格式将在另文中讨论。一般地,采用虚拟网页技术建设一个网站应分为三个设计阶段。

(1)素材准备阶段 是网站设计工作的基础,主要任务是通过网站应用分析,收集、整理、确定和建立素材库。以网上商场为例,素材准备阶段的主要任务就是确定所销售产品的相关信息,包括产品分类、产品说明的文字介绍、产品相关图片或动画宣传形式等内容,并制作成电子数据文件放入素材库中。

(2)中间件设计阶段 在素材库建立后,应根据网站各种可能的网页要求,确定可能用到的信息区块(称为中间件),并进行设计。所谓中间件可以认为是网页内容的组成部分,相当于 HTML 标记段,一般地中间件应是将来设计网页时需要平凡使用的内容,常是多个网页设计所需的公共部分,是相对完整和相对稳定

的内容,如网上商场情形,网页设计所需的“销售商品的完整文字介绍的形式组织”、“商品的图片显示及图片说明的形式组织”等均可设计为中间件。中间件的设计,采用在素材库上定义相应 HTML 标记段的扩展标记图来实现。中间件虽然从最终浏览显示效果看等同于 HTML 标记段,但一般地在素材库中并不真正存储该 HTML 标记段,而是相应的扩展标记图文件。

当然,当要求浏览速度时,可将扩展标记图和通过自动转换程序得到的相应 HTML 标记段同时存放在素材库中,指明它们的联系。这些 HTML 标记段可随时删除而不会影响相应虚拟网页的内容,而且当基础素材发生改变时,可通过相应扩展标记图作相应更新。

(3)虚拟网页设计阶段 这是网站设计的最后阶段。根据应用需要设计虚拟网页,与传统方式设计网页不同的是,不是编写实际的 HTML 文档,而是基于素材库来定义相应的扩展标记图,对应于完整 HTML 文档(而非 HTML 标记段)的扩展标记图又称为虚拟网页。

必要时,已定义的虚拟网页仍可作为中间件放入素材库中,以供定义其它虚拟网页时使用。只是自动转换程序在动态形成 HTML 文档时,应去掉这类中间件中多余的 HTML 标记,如<HTML>、<TITLE>和<BODY>等标记,以形成一个有效的 HTML 文档。

4 示例与实例化

4.1 示例

对如下 HTML 文档表示的网页(为叙述方便,HTML 文档中对相同标记加上标号来指明不同标记段),可以通过定义虚拟网页(扩展标记图) G_1 或 G_2 来表示。图1中,用()符号来表示节点,图中边是有序的,用数字来区分,用于指明标记段在 HTML 文档中的顺序关系。

```
<html>
<body>
<h1 align=center>成都旅游</h1>
<table>
<tr><td>峨眉山</td><td>...景点介绍1...</td></tr>
<tr><td>九寨沟</td><td>...景点介绍2...</td></tr>
</table>
<h1 align=center>重庆旅游</h1>
<table>
<tr><td>大足石刻</td><td>...景点介绍3...</td></tr>
<tr><td>红岩纪念馆</td><td>...景点介绍4...</td></tr>
</table>
</body>
</html>
```

这里,虚拟网页 G_1 是把 $h1^1$ 、 $table^1$ 、 $h1^2$ 和 $table^2$ 四个标记段看作四个节点,这四个标记段作为基础素材放入素材库中。而虚拟网页 G_2 则是把 $h1^1$ 和 $table^1$ 标记段(记 $h1^1-table^1$)看作一个节点而把 $h1^2$ 和 $table^2$ 标

(下转封四)

(上接第 82 页)

记段(记 $h1^1-table^1$)看作另一个节点。 G_2 的设计可以有两种方式,一是将 $h1^1-table^1$ 和 $h1^2-table^2$ 两个标记段作为基础素材放入素材库中,然后直接定义 G_2 。二是将 $h1^1$ 、 $table^1$ 、 $h1^2$ 和 $table^2$ 四个标记段看作基础素材放入素材库中,首先定义 $h1^1-table^1$ 和 $h1^2-table^2$ 两个标记段相应的中间件(扩展标记图)并放入素材库中,然后采用这两个中间件来定义 G_2 。

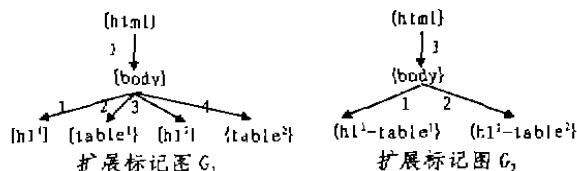


图 1

4.2 实例化

素材库中的中间件,一般只存储扩展标记图而无相应的 HTML 标记段,便于一致性维护。由于从扩展标记图自动动态形成 HTML 标记段会使效率降低,对那些平凡使用的中间件,必要时可以在素材库中同时临时存放相应的 HTML 标记段,称为扩展标记图的实例化。

同样地,从虚拟网页自动动态形成 HTML 文档(实网页)比直接浏览 HTML 文档的效率低,对那些平凡浏览的、或要求访问效率的虚拟网页,必要时也可以将动态形成 HTML 文档作为实际文档加以临时存放,称为虚拟网页的实例化。

当基础素材发生改变时,很容易通过相应的虚拟网页和扩展标记图对其实例化对象作相应更新,而且必要时可以随时删除这些实例化对象。

结束语 传统网页设计,网页常用一个实际的 HTML 文档(实网页)给出,即描述信息的组织结构又

包含具体内容。通过分析采用 HTML 文档设计网页存在的不足,本文提出通过建立基本素材库、引入扩展标记图、设计虚拟网页的新思想。

虚拟网页技术,将数据内容和结构分离,网页采用虚拟网页给出,仅描述信息的组织结构,数据内容统一放在素材库中,与传统网页设计相比,虚拟网页技术具有下述特点和研究意义:

(1) 素材库中同一内容可以同时用于多个虚拟网页,而当素材内容改变时,由虚拟网页动态形成的 HTML 文档内容会相应变化,有很好的数据共享性和一致性。

(2) 虚拟网页可以基于素材库中的中间件(扩展标记图)来定义,有利于网页模块化设计和提高设计效率。

(3) 便于适应不断变化的浏览要求,随时调整和增加网页。

(4) 虚拟网页是一种半结构数据模型,与标记图^[1]、OEM^[2]和 OIM^[3]等数据模型有着许多相似之处,可借鉴已有研究成果。

(5) 便于 Web 信息查询语言和查询优化的研究,可借鉴 TGQL^[1]、OIQL^[4]和 STRUQL^[5]等语言的研究方法和技术。

参考文献

- 1 陈澄,徐宏炳,王能斌. 基于标记图的 Web 数据模型. 计算机学报,1999,22(3):306~312
- 2 Papakonstantinou Y, Garcia-Molina H, Widom J. Object exchange across heterogeneous information sources. In: Proc CDE, Taipei: IEEE Computer Society, 1995: 251~260
- 3 王宁,徐宏炳,王能斌. 基于带根连通有向图的对象集成模型及代数. 软件学报,1998,9(12):894~898
- 4 McHugh J, Abiteboul S, Goldman R, et al. Lore: a database management system for semistructured data. ACM SIGMOD, 1997, 26(3): 54~65
- 5 Buneman P, Davidson S, Hillebrand G, et al. A query language and optimization techniques for unstructured data. In: Proc of SIGMOD'96, New York, 1996: 505~516

计算机科学

(1974年1月创刊)

第28卷第11期(月刊)

2001年11月25日出版

中国标准刊号: ISSN 1002-137X
CN50-1075/TP

定价: 12.00元 国外定价: 5美元

邮发代号: 78-68

发行范围: 国内外公开

主管单位: 国家科学技术部

主办单位: 国家科技部西南信息中心

编辑出版: 《计算机科学》杂志社

重庆市渝中区胜利路132号 邮政编码: 400013

电话: (023) 63500828 E-mail: jsj@swic.ac.cn

主 编: 朱宗元

印刷者: 重庆科情印务有限公司

总发行处: 重庆市邮政局

订购处: 全国各地邮政局

国外总发行: 中国国际图书贸易总公司(北京399信箱)

国外代号: 6210-MO