计算机科学2000Vol. 27№. 3

一个 Web 站点自动生成系统*)

A System of Web Sites Automatic Generating

肖庆文　林作铨

(汕头大学计算机科学研究所　汕头515063)　(北京大学信息科学系　北京100871)

Abstract　With the development of Internet/Intranet, automatic generating tools are needed to build lots of web sites. In the paper, a system of web sites automatic generating, which is based on knowledge base and non-monotonic reasoning, is put forward according to information construct on WWW, and then a sample using the system to generate a web site for a department is described.

Keywords　Web site, Knowledge base, Non-monotonic, Automatic generating

1　引言

目前,Internet/Intranet 正在飞速发展,而 Web 是 Internet/Intranet 上一种最有效的信息交流方式,用户可以通过访问 Web 站点来获得信息、服务,并反馈意见。因此,对于普遍的政府部门、企业、学校等单位来说,迫切需求建立 Web 站点,也需要一种能快速生成与维护 Web 站点的工具,然而,现存的 FrontPage 等 HTML 页面编写工具,自动化、智能化程度不高,操作复杂烦琐,要求用户具有较高的计算机技能,普通用户难于掌握。因此,给出一个 Web 站点自动生成工具,帮助企事业单位方便、快速地建立起 Web 站点是非常有意义的。

对于 Web 站点的自动生成,本文提出了一种使用人工智能技术、基于知识库、采用非单调推理的生成方法,并且根据这种思想实现了一个 Web 站点自动生成系统——WSAGS。WSAGS 能够协助用户完成建立 Web 站点的基本工作,用户只需要通过人机交互界面输入单位信息即可生成一个 Web 站点,使用 WSAGS 系统,Web 站点建立过程中的许多技术细节对用户透明,例如,编写 HTML 页面,在 Web 上发布数据库等。

2　WWW 上的信息结构

WWW 是在 Internet/Intranet 上组织和发布信息资源的一种方式,是目前全球最大的信息系统。WWW 是由可通过各种协议(主要是 HTTP 协议)获取的数字化文件组成的数据网络,它提供了一种获取 Internet 上不同资源的统一方式。它由许多称为 Web 页的超媒体文档组成,这些文档用 HTML 语言书写,包含多种媒体对象和指向其它文档的指针(超级链接)。Web 文档散布在世界各地的 Web 服务器上,每个服务器自主地管理自己的资源,没有统一的管理机制。浏览器使用 HTTP 协议实现 Web 文档的浏览。随着各种界面友好的浏览器(如:Netscape Navigator, Microsoft Internet Explorer)的出现,WWW 的信息得以迅猛发展。

WWW 上的信息用 Web 站点上的页面表示,页面文本按照 HTML 标记语言的格式来编写。这些页面可以通过 HTML 中的标记和各种多媒体资源构成链接,从而形成超级文本,同时,还可以通过标记链接到本站点或其它站点的页面,从而形成了数据网络。

WWW 上的各种页面资源一般都是通过统一资源定位标记 URL(Unified Resources Locator)来确定,页面中的各种多媒体资源也都是如此。每一个完整的 URL 包含四个部分:①获取该资源的协议;②该资源所在的 Internet 站点位置;③该资源在该站点的目录位置;④该资源的文件名。但是,WWW 上的页面中的 URL 一般都是不完整的,是相对于当前页面所在位置的相对路径。

用 HTML 语言写成的源文本(页面)由表示信息的文本块和起控制作用的标记矢量(Tags Vector)组成;而标记矢量中的每一个元素又是一个包含各种属性的复杂特性集,这些属性大致可以分为两类:页面中各种资源的 URL 和页面元素的形态属性[7]。因此,当固定标记矢量而改变信息内容时,可以生成外观相似而内容不同的页面。

*)本文获得国家自然科学基金和国家863计划资助

对于同一类型的单位，它们要发布的信息实际内容不同，但是抽象出来的信息框架却是相同的，例如，某个大学的化学系与物理系，都要发布系介绍、专业介绍、教师介绍……等等，这些页面的彼此关系（链接）也可以是相同的，把这些信息框架知识存储在知识库中，当某个此类单位要生成 Web 站点时，就可以根据信息框架知识询问信息实际内容，自动生成各种 Web 站点页面。

3 WSAGS 系统模型

对于一般用户来说，熟悉各种 HTML 标记及其用法是比较困难的，但对 Web 站点上发布的单位信息却非常熟悉。因此，Web 站点自动生成系统（WSAGS）的研制目标是：生成 WWW 上的信息载体——HTML 文档页面时，只需要一般用户提供将在 Web 站点上发布的企事业单位文字和图片信息，不需要用户知道各种 HTML 标记及其用法，也不需要用户管理页面之间的链接以及页面内部多媒体资源的链接。在这个生成过程中，HTML 文档的编写以及各种服务应用程序的编程等技术细节对用户是透明的，用户只要回答系统提出的问题即可。

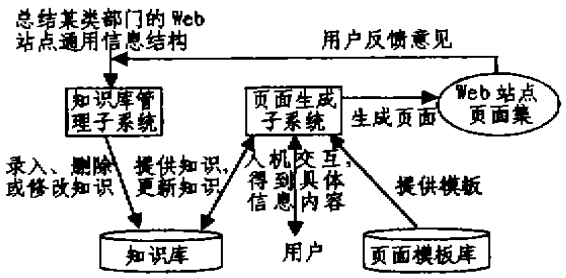

图1 系统模型图

WSAGS 系统的模型如图1所示，现说明如下：

a）专家总结行业 X 中需要在 Web 上发布的内容，经过抽象得到 X 类型的 Web 站点通用信息结构；

b）录入人员通过知识库管理子系统添加（或删除、更改）知识库中的信息框架知识、页面间链接知识等；

c）对于行业 X 中的每个单位 x_i，在建立 Web 站点时，启动页面生成子系统。页面生成子系统根据系统知识库中储存的知识，通过问答式人机交互，得到 x_i 的文字与图片等具体信息内容，然后应用用户选择的页面模板生成 Web 站点页面，添加到 Web 服务器上，一个 Web 站点就初步建立了；

d）用户在使用系统中，发现需要添加、删除或更新某些知识，再重复 b，c 步骤。

知识库用于存储向用户提问及逻辑推理时所用到的知识。一个行业的所有单位具有相同的信息框架，知识的信息框架改变时可适应不同的行业，用户也可以部分更改知识库中的知识使之适合自己的特殊需要。知识库中的知识采用树型结构存放，根节点是 Web 站点的首页（HOMEPAGE）的知识，孩子节点为父节点的后继页的知识。节点信息用下述数据结构表示：

```
Class node
{
    int      ID; //知识编号
    char*    Node_Name; //节点主题描述
    int      Node_Type; //节点类型(一般页面,服务程
    序页面)
    char*    Node_Question; //询问用户的问题
    node*    Node_Father; //父节点指针
    int      Node_Son_Num; //后代个数,-1表示后代个
    数不能预先确定
    node*    Node_Son[]; //子节点指针集
}
```

页面模板库中的模板为预先设计好的若干页面，具有不同的浏览器显示风格。页面模板主要定义：页面背景颜色，标题字体（字型、字号、字体颜色），正文字体等等。为了使得生成的站点具有个性化，WSAGS 预先准备了较丰富的页面模板，供用户选择。

WSAGS 中的知识推理采用非单调的推理思想[3]。假设 P 无矛盾，如果根据知识集合 P 中的知识 p_i 能生成页面 s_i，则可以把 s_i 加入到 Web 站点页面集合 W；引入新知识 q，q 与知识 p_i 无矛盾，则 s_i 保留；否则，页面 s_i 从 W 中删除，如果 q 与 P 中任何知识无矛盾且能生成页面 t，则 t 也加入到这个集合中。

算法：

1. 生成知识集合 P 置空
2. 页面集合 W 置空
3. 根据知识编号，从知识库中获取下一知识 q
4. flag=True
5. 从 P 中取一知识 p_i
6. 如果 p_i 与 q 矛盾，则 flag=Flase，从 W 中删除与 p_i 对应的 s_i
7. 如果 P 中还有知识未处理，转5
8. 如果 flag=Flase，转10
9. 根据 q 生成页面 t，q 添加到 P 中，t 添加到 W 中
10. 如果知识库中有知识未处理，转3

对于不同于静态页面的动态页面，需要服务器端 Web 应用程序的支持。服务器端应用程序一般是 CGI、ASP 或 API 动态交互程序。WSAGS 中内置多种 Web 服务器端网关接口通用程序及其实例化生成程序，例如，系统中有一个通用的 Web 数据库查询程序，当应用于特定数据库时，Web 端数据库服务生成程序生成数据库查询页面，更改查询程序中的 SQL 语句的

知识，得到特定数据库的查询程序。系统还提供外部程序模块接口，可以添加用户编写的特定 Web 服务程序及其生成程序。

4 Web 站点的生成

下面使用 WSAGS 模拟生成一个计算机系的 Web 站点。首先，进行 Web 站点自动生成的准备工作：总结系级 Web 站点的信息框架，生成关于系级站点的知识库，知识库一经建立，可以反复使用。一个简单的系级的 Web 站点信息框架如下：

◇首页，包括的信息内容：本站点介绍，最新消息。

◇系介绍，介绍本系的历史、现状以及未来规划和展望。

◇教师介绍，介绍本系的师资力量、科研力量，并提供人员查询、教师主页。

◇专业介绍，介绍本系各专业情况，包括专业介绍、分配前景等。

◇学习条件，介绍本系的勤工俭学，奖学金等情况，并提供学生个人主页空间。

◇布告栏，系发布的一些公告。

这些信息用树状结构图表示如图2。

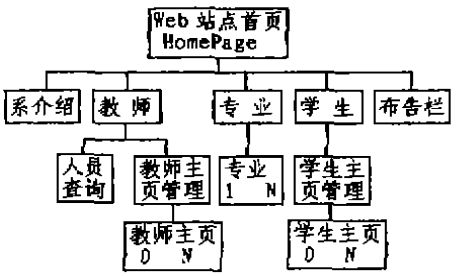

图2 系 Web 站点结构图

使用知识库管理子系统，把上述知识添加到知识库中。在知识库中，节点的内容如表1所示。

然后，使用 WSAGS 生成一个计算机系的 Web 站点。

页面生成子系统查询知识库，得到页面问题，询问用户；用户输入相应问题的答案并指出添加的图片位置；系统生成页面、页面上的超文本链接按照知识库提供的节点父子关系设置。例如，生成"HOMEPAGE"页面时，系统提问"本站点介绍"，用户输入"这是 S 大学的计算机系站点，欢迎光临！"，提供一幅图片，系统按照这些信息生成页面，并设置到系介绍、教学员工等页面的链接。

表1 知识库内容表

ID	页面主题	页面类型	本页问题	父节点	子节点个数	子节点
1	首页	NORMAL PAGE	本站点介绍	0	5	2、3、4、5、6
2	系介绍	NORMAL PAGE	本系历史 现状 未来发展展望	1	0	
3	教师	NORMAL PAGE	教学情况 科研情况	1	2	7、8
…	…	…	…	…	…	…
7	人员查询	DATABASE MANAGE SERVICEPAGE		3	0	
…	…	…	…	…	…	…

当生成动态页面即页面类型不是"NORMAL-PAGE"的页面时，系统调用特定的生成模块。例如，生成"人员查询"服务，WSAGS 调用数据库服务生成模块，根据用户提供的"人员数据库"的位置，得到数据库中的所有数据域；然后用户从中选择按姓名、教授课程查询，系统生成查询页面，并把查询语句中的数据域改为姓名、教授课程，实例化通用查询模块，得到人员查询网关接口程序。

如上，系统逐个生成所有页面，添加到 WWW 服务器上，就生成了 Web 站点。

参 考 文 献

1 Quek C Y, Mitchell T. Classification of World Wide Web Documents. School of Computer Science Carnegie Mellon University, thesis, 1997, 5

2 梁锴，林作铨. 基于 Windows NT 的 Intranet 的研究与实现. 见：中国智能自动化学术会议(CIAC' 98)论文集. 1998. 5

3 Hypertext Markup Language. Available at: http://www.w3.org/TR/REC-html40/

4 徐东晖，蔡希尧. 一种新型的基于 WWW 的应用开发平台. 计算机科学，1998，25(3)：70～73

5 Intelligent Agents start. Available at: http://activist.gpl.ibm.com:80/whitePaper/ptc2.htm

6 政府上网工程. Available at: http://www.gov.cn/

7 沈达阳，林作铨. Internet 信息收集 Agent 及其搜索方法. 计算机系统应用，1998，5：18～21

8 陆汝钤. 人工智能. 科学出版社，1996. 9