超文本结构化转换算法的研究与实现*

郑庆华, 由渊霞, 袁文斌

(西安交通大学 计算机科学与工程系,陕西 西安 710049)

E-mail: qhzheng@xjtu.edu.cn

http://www.xjtu.edu.cn

摘要: 超文本是一种非结构化的文档,它虽然不支持跨页查询和全文检索,但却是Internet上信息组织与存储的重要方式.提出了一种将超文本转换为结构化数据库的算法.分析了超文本结构化转换的需求,运用图论分析并描述了超文本的转换模型与实现算法.该算法在鲁迅数字图书馆系统中得到了实际应用和验证.

关键词: 数字图书馆;超文本;结构化;数据库

中图法分类号: TP311 **文献标识码**: A

超文本是采用HTML语言编写或制作的非结构化或半结构化文档,它与HTTP和URL一起成为WWW中的三大关键技术.超文本由于具有平台无关、支持基于内容的联想式超链接信息组织方式以及多媒体化的人机界面,因此成为Internet上信息组织、存储与发布的主要方式之一[1].但是,由于超文本是一种非结构化文档,一般仅适合于信息的浏览和导航.文献[2]指出了超文本系统在支持信息检索方面存在的3点不足:(1) 严重的迷路问题;(2) 无法支持对信息的直接定位;(3) 查找信息的效率太低.目前,ISP(internet service provider)站点建设的主题已经从接入服务转向内容服务,如何有效地管理门类繁多的超文本信息,并能以快捷、有效和有价值的服务吸引用户,这是一个十分重要的问题.文献[3,4]提出了解决此类问题的若干途径,如,建立丰富的导航、浏览图、主航线,提供查询机制等等.查询当然是最有效的方法,但不仅要查询内容,而且还要查询信息的组织结构.

目前,实现超文本或主页信息的跨页查询,一般有两种解决途径.(1) 采用WWW服务器自带的索引服务器,如Windows NT Server平台下与Web Server IIS 4.0相配套的Microsoft Index Server.这种方法只能实现字符串匹配查询,而无法实现按主题查询,而且检索效率低下,同时受平台的限制,其可移植性较差[3].(2) 把超文本和数据库相结合,利用数据库强大的数据组织、管理、查询能力来提高整个超文本系统的性能.不过,这方面的研究大多局限于通过数据库查询,将查询结果动态地填充到主页框架以解决主页和数据库间的连接问题.但这样对超文本的表现能力,包括超链接、多媒体、动态特性和整体效果等产生了很大的限制.

本文提出了一种解决上述问题的新思路:通过将非结构化的超文本文件集自动转换成结构化的数据库,并对数据库中的超文本记录的特征字段进行标引(人工或自动均可),形成完整的超文本

* 收稿日期: 2000-04-13; 修改日期: 2000-07-26

基金项目: 国家863高科技发展计划资助项目(863-317-01-04-99);机械制造系统工程国家重点实验室基金资助项目;西安交通大学科学研究基金资助项目(98-007)

作者简介: 郑庆华(1969-),男,浙江绍兴人,博士,副教授,主要研究领域为CSCW,多媒体远程教育技术;由渊霞(1977-),女,山东青岛人,硕士生,主要研究领域为多媒体远程教育;袁文斌(1975-),男,江苏江阴人,助教,主要研究领域为多媒体远程教育.

数据库,在此基础上开发相应的基于 Web 的按主题和内容等的检索引擎(如公共网关接口 CGI 程序),实现对超文本或主页进行按主题、关键词、内容等的各种查询.此外,这种方法还可以实现对超文本文件之间链接关系的快速分析和对超文本文件集合的有效管理.

本文在分析超文本文件转换成数据库所要达到的基本目标和需求的基础上,提出了一种超文本结构化转换的模型及相应的实现算法,并在数字图书馆系统中得到实际应用和验证.应用表明,本方法可以较好地解决超文本的高效检索和管理问题.

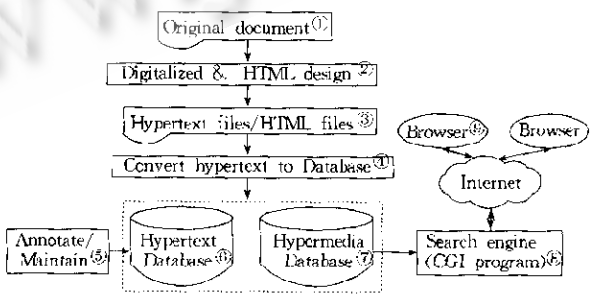
1 转换需求与模型

将超文本文件集转换成超文本数据库,应满足以下基本需求:

- 无损性:转换过程不能丢失或放弃原超文本的正文信息和描述信息;
- 可还原性:当需要时,可以从超文本数据库中还原出原来的超文本文件;
- 链接关系自动分析:由系统自动分析超文本之间的链接关系;
- 链接关系显式表示:将这种链接关系显式地存储在数据库的链接关系字段中,并且可以用树状或网状的形式反映给用户.

其中,无损性和可还原性保证了超文本文件集和转换后的超文本数据库在内容上的相互等价,链接关系的自动分析和显式表示则反映了超文本的组织结构.在此基础上,为了实现对超文本数据库的有效管理,并支持基于 Web 的各种信息查询,还应提供以下功能:

- 特征标引:允许工作人员对导入后的超文本记录进行特征标引,如关键词、主题词、摘要、作者、日期等,因为这些信息都是 Web 用户查询所需要的.但转换程序无法从超文本文件中直接获取这些信息.
- 安全删除:当删除超文本数据库中的一条记录时,系统将自动判断该记录所对应的超文本是否与其他记录所对应的超文本相关联.若存在关联,则给出警告.并支持以下两种方式的删除:一种是仅删除当前记录,另一种是删除与此记录相关的全部超文档记录.
- 安全替换:实现新的超文本文件对原文件的替换,并自动更新与此超文本相关的链接关系.
- 链接关系查看:以树状或网状形式反映出某一超文本与其他超文本文件之间的链接关系,使工作人员可以一览全局.
- 基于 Web 的主题或内容查询:通过开发相应的 CGI 程序,支持 Web 用户对超文本数据库进行按主题词、关键词、作者、摘要、日期等信息的查询.



①原始文献,②数字化及主页设计,③超文本集/HTML文件,④超文本结构化转换,⑤标引/维护,⑥超文本数据库,⑦超媒体目录数据库,⑧检索引擎(CGI程序),⑨Web用户.

Fig. 1 The model of converting unstructured hypertext to structured database

图1 超文本结构化转换模型

根据上述需求,我们设计了如图 1 所示的超文本结构化转换的实现模型.

总体上,本模型可以分为以下 3 部分:

• 超文本结构化转换模块:在用户将原始文献资料数字化并制作成主页(HTML 文件)的基础上,通过本模块将其转换成超文本数据库,并保证转换的无损性、可还原性和链接关系自动分析;

• 标引和维护模块:对转换后的超文本数据库进行特征标引,以形成可供 Web 用户查询或检索的完整的超文本数据库,并提供备份、恢复、记录的安全删除、替换等维护操作;

• 检索引擎模块:这是一个 CGI 程序,支持 Web 用户对超文本数据库进行按主题、特征词、内容等信息的查询.

限于篇幅,本文主要阐述超文本结构化转换问题的实现算法.

2 超文本结构化转换算法

2.1 超文本描述

超文本结构化转换算法(简称 HtoDB 算法)的求解核心是:

(1) 如何完整地分析超文本文件之间的链接关系,并将这种链接关系由 HTML 语言的嵌入式隐式表示转换为超文本数据库中的显式存储表示;

(2) 根据超文本的特征(主要体现为 HTML 语法),自动提取其中的内容并写入超文本数据库的对应字段中,以实现一一转换,并保证无损性和可还原性.

为了解决上述问题,我们先对超文本进行必要的描述.

超文本结构可以用有向图 $G=(N,E)$ 表示[5,6]. 其中,N 表示超文本结点的集合,$N=\{n_1, n_2,\ldots,n_M\}$,$M=\|N\|$,每一个结点 n_i 表示一个超文本文件或媒体文件. E 表示有向边的集合,且 $E\subset N\times N$. 图 G 具有以下性质:

(1) 连通性. 设 S 为图 G 的入口结点(一般为 Index.htm 或 Default.htm 文件对应的根结点),则对于 $\forall\ n_i\in N$,至少存在一条从 S 到 n_i 的路径.

(2) 设 $\mathrm{In}(n_i)$ 表示链接指向结点 n_i 的结点数,即 n_i 的入度;$\mathrm{Out}(n_i)$ 表示由 n_i 指向其他结点的链接数,即 n_i 的出度,则除了入口结点 S 的入度为 0,即 $\mathrm{In}(S)=0$ 以外,其余结点的入度均大于 0,而所有叶结点的出度均为 0.

(3) 在图 G 中,所有结点总的出度与总的入度相等.

$$\sum_{i=1}^{M}\mathrm{In}(n_i)=\sum_{j=1}^{M}\mathrm{Out}(n_j). \tag{1}$$

进一步分析可以得出:虽然超文本有向图 G 中的链接关系十分复杂,但归根到底是由每一结点的前趋和后继关系组成的,只要分析得出每一结点的前趋和后继关系,进而就可以得出整个有向图 G 的结构. 因此,链接关系分析的实质是:对图 G 中的每个结点搜索出其前趋结点集和后继结点集. 即对 $\forall\ n_i\in N$,确定以下两个集合:

$$D(n_i)=\{n_j\mid\forall\ n_j\in N,(n_j,n_i)\in E\}, \tag{2}$$

$$S(n_i)=\{n_j\mid\forall\ n_j\in N,(n_i,n_j)\in E\}. \tag{3}$$

其中 $D(n_i)$ 表示结点 n_i 的所有前趋结点集(又称上链接结点集),$S(n_i)$ 表示结点 n_i 的所有后继结点集(又称下链接结点集),如图 2 所示. 且有

$$\mathrm{In}(n_i)=\|D(n_i)\|, \tag{4}$$

$$\mathrm{Out}(n_i)=\|S(n_i)\|. \tag{5}$$

式(2)和式(3)可作为超文本结点间链接关系分析的基本依据,而式(1)、式(4)和式(5)则可成为链接关系分析正确性判断的依据之一.关于链接关系分析正确性的测试,可参见文献[6].

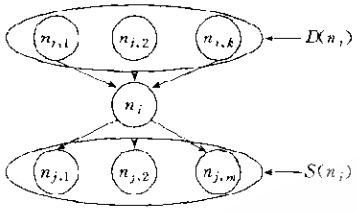


Fig. 2 The success and prefer hypertext node set of n_i
图2 超文本结点 n_i 的前趋结点集和后继结点集

超文本链接关系分析的另一个重要问题是如何保证完整性.关于这一点,有以下结点覆盖准则[6].

结点覆盖准则:超文本链接关系分析是结点充分的,即对于$\forall n_i \in N$,只要获得了 $D(n_i)$和 $S(n_i)$,即可保证链接分析的完整性.

2.2 超文本结构化转换的数据结构

超文本经结构化转换以后,以数据库方式保存了所有主页的内容及其附属多媒体文件的描述信息,数据库记录和超文本间形成了一一对应的关系,即超文本数据库中的一条记录对应于一个超文本文件的内容,同时,在相应字段中还保存了与此超文本结点对应的链接关系字段:前趋结点集 $D(n_i)$和后继结点集 $S(n_i)$.超文本结构化后的数据结构如下:

〈超文本记录〉::=〈主页编号,PK〉,〈HTML 文件路径〉,〈上链接关系,$D(n_i)$〉,〈下链接关系,$S(n_i)$〉,〈主题〉,〈关键词〉,〈主题词〉,〈摘要信息〉,〈主页正文信息〉,〈设计者〉,〈设计时间〉,〈设计备忘〉,〈删除标记〉.

其中,主页编号为超文本记录的关键字,由系统按类自动产生.上链接关系和下链接关系分别为该超文本文件所对应结点的前趋结点集 $D(n_i)$和后继结点集 $S(n_i)$,由系统自动分析所获得.关键词、主题词、设计备忘、主页摘要等字段信息由管理员录入标引,其他字段由系统自动分析超文本文件生成.

值得指出的是,通过基于 Web 的检索引擎,Internet 用户可对超文本数据库中的关键词、主题词、主页摘要、主页正文、设计者、设计时间等信息进行查询,而上链接关系和下链接关系字段则使用户可以直观地获得与本超文本结点相关的前趋与后继结点.

2.3 算法设计与实现

根据上述超文本及其数据结构的描述,我们设计了以下超文本结构化转换算法 HtoDB.

算法. 超文本结构化转换 HtoDB 算法

输入:① 超文本文件所在的存储目录;② 主文件(如 Index.htm).

输出:超文本数据库.

功能:实现超文本文件集向超文本数据库的结构化转换.

实现步骤:

Step 1. 初始化; /* 建立超文本标记的模板库(模板库记录了 HTML 文件中的检索标记,如"href"等),设置路径和根结点对应的文件名 */

Step 2. 调用 ReadIn()函数读入主文件 Index.htm,进行根结点分析,建立超文本数据库的首

记录; /* ReadIn()为超文本文件读入函数 */

Step 3. 获取超文本文件的个数,设为 N,并假设超文本文件集合为 HT={Index.htm,HF[1],HF[2],...,HF[N]};

Step 4. 对于每个超文本文件,执行 Step 5~7;

Step 5. 调用主页编号生成函数 GetLeastAvailNum()获取当前主页的编号,作为该超文本记录的主关键字;

Step 6. 调用 Link-Analyze 过程,得到该超文本文件对应结点的所有后继结点的集合 $S(n_i)$以及所有前趋结点的集合 $D(n_i)$;

Step 7. 在超文本数据库中创建本结点对应的记录,并将主页编号、HTML 文件路径、主题、主页的正文信息、设计者、设计时间等字段设置为默认值,将下链接关系字段设置为 $S(n_i)$值,将上链接关系字段设置为 $D(n_i)$值;

Step 8. 如果是最后一个超文本文件,则转 Step 9,否则转 Step 4;

Step 9. 关闭数据库,转换结束.

Link-Analyze 过程根据超文本标记模板库中的特征值对超文本文件中的内容进行扫描,以生成后继结点集合 $S(n_i)$;对前趋结点集合 $D(n_i)$的生成,则需要通过扫描其他超文本文件得到.假设超文本文件的个数为 n,则转换算法的复杂度为 $O(n)$,即一次扫描便可实现全部转换.

3 实际应用与结论

基于以上模型及算法,我们在"鲁迅图书馆"数字化网络信息系统*和中国文献保障体系(China Academic Library and Information System,简称 CALIS)西北地区文献中心——西安交通大学"钱学森图书馆"数字化图书馆**建设中得到了应用验证.应用由我们开发的超文本结构化转换及其管理系统以及基于 Web 的检索引擎,在"鲁迅图书馆"建立了"绍兴古桥"、"地方戏剧"和"绍兴黄酒"等 8 个特色文化数据库,实现了超文本到数据库的结构化转换、超文本间链接关系的分析和增、删、改等维护功能,并向 Internet 用户提供了对数字化文献基于 Web 的、按主题和内容方式的检索功能.我们在"钱学森图书馆"特色文化数据库的建设中,建立了"钱学森"生平事迹全文数据库和自动化学科文献数据库,记录数据接近 10 万条.

随着信息化过程在许多领域的深入发展,弥补超文本先天不足的需求正在变得日益强烈和紧迫.传统的情报机构要在信息化的道路上发展,就不能只停留在把主页发布到 Internet 上的水平,而应站在信息智能服务的高度,为用户提供全面、便利的信息获取手段[4,7,8].本文在这方面做了有益的探索.实际应用表明,利用本算法能有效地将非结构化的超文本转换成结构化的数据库,并能自动地分析超文本之间的链接关系,从而实现对超文本文件的有效管理,并支持用户对超文本进行以主题、关键词、摘要、正文等内容的全文或跨页查询.

References:

[1] Fox, E. A. Networked digital library of thesis and dissertation. D-Lib Magazine, 1998,2(2):20~27.

[2] Zhao, Dan-qun. A discussion of the hypertext system's information retrieval efficiency. Information Theory and Practice, 1999,22(2):90~92 (in Chinese).

* 横向合作课题,于 2000 年元月通过浙江省科委的技术鉴定.http://library.zjonline.com

** 教育部 CALIS 项目,于 2000 年 11 月通过验收.http://202.117.24.50

[3] Mitchell, S., Mooney, M. INFOMINE——a model Web-based academic virtual library. Information Technology and Libraries, 1996,2(3):20~25.

[4] Yang, Ji-guo, Yang, Dong-qing, Tang, Shi-wei. Approaches to integrate database and hypertext system. Computer Applications, 1997,17(5):32~36 (in Chinese).

[5] Gonzalez, R. Hypermedia data modeling, coding, and semiotics. Proceedings of the IEEE, 1997,85(7):1111~1137.

[6] Jin, Ling-zi, Zhu, Hong. An introduction on sufficiency rule of testing of hypertext application software. Journal of Software, 1997,8:130~136 (in Chinese).

[7] Lang, Song-zhen, Zhu, Xiao-feng, Li, Xue. ISP——the necessary choice of our information institutions under the new environment. Information Theory and Practice, 1999,22(2):78~80 (in Chinese).

[8] Ye, Xiao-xin. Content service: the essential of domestic ISP's management. Information Theory and Practice, 1999,22(5):370~372 (in Chinese).

附中文参考文献:

[2] 赵丹群.超文本系统的检索能力评析.情报理论与实践,1999,22(2):90~92.

[4] 杨继国,杨冬青,唐世渭.数据库与超文本系统的连接.计算机应用,1997.17(5):32~36.

[6] 金凌紫,朱鸿.超文本应用软件测试充分性准则初探.软件学报,1997,8:130~136.

[7] 郎诵真,朱晓峰,李雪.ISP——我国情报机构在新形势下的必然选择.情报理论与实践,1999,22(2):78~80.

[8] 叶笑欣.内容服务:国内ISP经营的真正主题.情报理论与实践,1999,22(5):370~372.

A Practical Algorithm for Converting Unstructured Hypertext to Structured Database*

ZHENG Qing-hua, YOU Yuan-xia, YUAN Wen-bin

(*Department of Computer Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China*)

E-mail: qhzheng@xjtu.edu.cn

http://www.xjtu.edu.cn

Abstract: Hypertext is a kind of unstructured document. It is impossible to realize the search based on content and topic for hypertext documents. However, hypertext is one of the most important ways of information storage and organization in the Internet. Therefore, in order to realize the effective management and the search of hypertext documents, a new and practical method named HtoDB for converting unstructured hypertext to database is presented. In the paper, the requirements and functions for converting hypertext to database are analyzed, the converting model and algorithm are also put forward according to the graph theory. The algorithm and model presented in this paper are verified in the project of "LU XUN digital library system".

Key words: digital library; hypertext; structured database

* Received April 13, 2000; accepted July 26, 2000

Supported by the National High Technology Development Program of China under Grant No. 863 317-01-04-99; the Foundation of State Key Laboratory of Mechanical Manufacturing and System Engineering; the Science Foundation of Xi'an Jiaotong University under Grant No. 98-007