网络半结构化信息资源的描述

黄 奇 李 伟 接晓莉

(南京大学信息管理系 南京 210093)

[摘要] 要对网络信息资源进行更好的管理和查询,首先要建立一种合理的信息资源描述机制。metadata是描述网络信息资源的有力工具,但新的信息描述机制——linking机制不仅能表述metadata的内容,而且可以表达比metadata更丰富的语义,弥补metadata自身不能克服的一些缺陷。

[关键词] 半结构化信息 信息资源描述 metadata linking

[分类号] G203 A

A Study on the Description of Web Semistructure Information Resource

Huang Qi Li Wei Jie Xiaoli

(Information Management Department, Nanjing University, Nanjing)

[**Abstract**] To manage and query Web information, the first step is to set up a good mechanism for describing the information. This paper introduces what semistructure information is, focusing on how to describe this kind of information. Metadata is considered to be a powerful mechanism for describing Web information. In this paper, the advantages and disadvantages of metadata are analyzed, then a kind of new information description mechanism—linking mechanism is presented. Linking can not only define metadata but also be more expressive than metadata. It can make up some implicit disadvantages of metadata.

[**Keywords**] semistructure information information description metadata linking

随着数字化、网络化技术的飞速发展,数字图书馆建设成为图书情报界重要的研究方向,其中网络信息资源组织是数字图书馆建设的核心内容。网络信息资源的动态性、分布性、多元性和无序性等特点,使信息的查找和检索变得越来越困难。对于21世纪的信息用户和信息管理者来说,困扰他们的不是信息太少,而是信息过多的问题。因此,如何对网络信息资源进行合理的描述,组织、序化网络信息资源,提高信息利用率,是当前重要的研究课题。

1 网络半结构化信息资源

目前,网络半结构化数据日趋丰富。完全结构化数据有非常良好的数据结构,如关系数据库、面向对象数据库中的数据。完全无结构数据是指声音、图像文件等无模式数据。而半结构化数据是介于完全结构化数据和无结构数据之间的一种数据类型。半结构化数据虽然有一定的结构,但却是不严格的、多变的和不完整的。

从网络的信息层次来看,网络半结构化信息的研究对象分为3个层面:网页层面、网站层面、网络层面。

——WWW网页:最主要的研究方向。

● HTML(Hypertext Markup Language)——由于其在目前网络资源描述格式中所占的比例最高,所以有关研究特别多。

● XML(eXtensible Markup Language)——作为一种新的网上数据交换的标准,正在引起人们极大

收稿日期:2001—10—20

的关注。XML是标准的通用标记语言SGML[ISO 8879]的一个子集，用于支持Internet上有结构文档的交换。和HTML相比，XML是面向内容的，它具有更多样化的结构和更丰富的语义，并具有可扩展性良好、易于掌握、自描述等特点，适用于Web上的数据交换。可以预言，XML将成为数据组织和交换的事实标准，大量的XML数据将出现在Web上。XML数据模型与半结构数据模型有着很多的相似性，即它既为半结构数据的研究提供了广阔的应用前景，同时也推动了半结构化数据研究的发展。

——网站的半结构化研究：充分利用网页内容、锚文本、网页链接、链接的兄弟关系等进行导航。

——网络的半结构化研究：通过挖掘利用网络信息半结构化的特点，设计智能搜索引擎，提供某一主题的高效检索。

要更好地组织网络半结构化信息，首先要对半结构化信息进行合理的组织描述。

2 半结构化信息资源的描述

目前，大多采用带标记的有向图作为半结构化数据模型，最典型的就是OEM(对象交换模型)模型。概括地说，主要有两种描述方法：

2.1 基于逻辑的描述形式

在已经提出的半结构化数据模式的描述形式中，基于逻辑的描述形式是重要的一类，如一阶逻辑(first—order logic)、描述逻辑(description logic)以及Datalog等。它们非常类似，但在表达能力等方面有所差别，其中比较典型的是基于Datalog的模式描述形式。

2.2 基于图的描述形式

由于半结构化数据一般采用带标记的有向图来表示，所以这种描述形式的一个显著优点是模式和数据采用同一种数据模型(图模型)，非常便于处理。模式图通常是一个有根、边上带标记的有向图，其边上的标记可以与数据图相同，也可以加以扩充，如允许类似于"name | address"的形式，或采用特定形式的规则(如一元谓词)，等等。对模式图中的节点，可以加以一定的注释，表明其代表的语义或其它特定的含义，其中最有代表性的是OEM。

此外，还有概念模型。通过一个自然简单的方法，了解HTML页面的内部结构。它不同于OEM，而类似于人对文档的概念化。它提供虽然很少却十分有效的高层结构，用于描述文本的内容(如通过引入LIST对象解决了图、树描述方法所不能解决的LIST表问题)。另有一套相应的规则，把内容自动映射到概念模型中。但到目前为止，还没有相应的查询语言。

2.3 半结构化模式的特点

- 先有数据，后有模式。一般是先进行查询，查询结果即为数据结构及其模式。
- 用于描述数据的结构信息，而不是对数据结构进行强制性约束。
- 规模可能很大，甚至超过源数据的规模，而且因数据的不断更新而处于动态的变化过程之中。
- 不讲求精确性，可能描述其中一部分结构，也可能根据数据处理的不同阶段的视角而不同。
- 非常灵活，能满足网络这种复杂分布式环境的要求。
- 加大了数据处理的难度。

3 从metadata到linking组织描述

为了描述网页半结构化信息资源，人们提出了元数据(metadata)的概念。提出metadata的目的，是将图书情报领域的分类法和标引技术普及到一般的网页制作者，以组织庞大的网络信息资源。metadata系统被认为是一个用于抽取构成对象的属性和方便信息访问的强有力的通用机制。

较早出现的元数据格式是MARC(主要被用来详细著录书目)，它是全球范围内公认的较为成熟的传统机读编目格式，其结构严谨，类目复杂；系统完善，但是并不适合对一般网络信息资源的描述。首先，网络信息资源描述格式并不需要那样复杂；其次，网络信息资源浩如烟海，让编目人员对每个网页都进行详细著录，需耗费相当的人力物力，这是不合实际的。

为了研究一种适用于描述一般网络信息资源的元数据标准，制定一种通用的网络著录规则，1995年3月由OCLC与NCSA联合发起，52位来自图书馆界和电脑网络界的专家共同研究产生了都柏林核心元数据集(Dublin Core，简称DC)。DC适合揭示各类型电子文献的内容和其它特性，能有效地对网上资源进行组织、分类、索引。

DC由15个基本元素组成，分成三大部分：①内容描述部分有题名、主题、说明、来源、语种、关联和覆盖范围；②知识产权部分有创建者、出版者、其他责任者和权限；③外形描述部分有日期、类型、形式和标识符。

DC比较易于应用到网络信息资源的描述中，著录数据与著录对象可以存在于同一文件中，也可以存在于不同文件中。

3.1 DC的优点

● 结构简单。数据元素的含义，易学易记，非编目人员也能很快理解。对网络资源的描述性编目，主要由资源制作者在制作资源的同时提供，这不仅降低了记录的制作成本，又能适应网络信息资源巨量增长的需要。在资源制作者描述的基础上，信息工作者则把主要精力放在对质量较高、稳定性较好的网络资源的标引和规范控制上，为用户构建高效实用的检索系统。

● 可重复性。DC规定所有元素都是可重复的，因而解决了多著者或多版本等重复元素的著录问题。

● 可选择性。著录项目可以简化，只须确保最低限度的7个元素（题名、出版者、形式、类型、标识符、日期和主题）即可。

● 可扩展性。各个DC地方版可以在15个元素的基础上增加新的元素或新的修饰词。允许资料以地区性规范出现，并保持元数据的一些特性，以便日后有扩充的余地。

● 可以与其它元数据连接使用，以弥补其自身的不足。在统一资源描述框架（RDF）下，可以实现与其它元数据的连接。

3.2 metadata存在的两大理由

● 提供在一个系统内扩张对象的通用机制。要注意的是，这一对象不一定是文献。

● metadata在系统中可以用于分组、排序并访问对象，即提供信息检索服务。

3.3 metadata的缺陷

● 主要是对文献的外部特征进行描述，虽然采用了主题这个元素对文献内部特征进行揭示，但描述得不够详细。

● 没有充分利用半结构数据中的结构信息。

● metadata机制看不出所描述的对象类型。

● metadata值对（名字和值）是不对称的，单向可读，域和域值角色不能交换。

实践中，有的研究者把DC和全文检索技术结合起来建立搜索引擎，但尚无实质性进展。故此，更切合对半结构化数据进行描述的机制——联接（linking）机制应运而生。

联接（linking）机制和metadata之间具有相似性，即它同样有能力抽取对象的属性，并给用户提供更有效的信息访问手段。

首先，metadata可以用linking的形式表示出来。

把metadata值对视为系统内可以识别的子对象，这样，metadata值对就成了对象的子对象，通过名字在锚（anchor）和metadata值对间建立映射。

示例1：

<doc1，<author，hq>>：其中<author，hq>为metadata值对，用p表示，有<doc1，p>，p视为doc的子对象，从而与锚的表达一致。

其次，linking机制可以表达比metadata更丰富的语义，能揭示关系对的类型，并且是对称的。

示例2：

<<"doc2"，"document">，<"hq"，"author">，"DocumentAuthoredBy">

<<"hq"，"supervisor">，<"lj"，"supervises">，"supervises">

第一句的语义为：doc2 is the document authored by hq。第二句的语义为：hq supervises lj。

通过这个例子可以发现，linking机制揭示了metadata模型不可能揭示的关系对的类型：document和supervisor。并且，它使metadata对称化了，其中hq这个link中的"锚"既可以做第一个关系对doc2的域值，也可以做另一个关系对hq的域名。

分析表明，linking机制更适合对网络半结构化信息的描述。在对网络半结构化信息资源合理描述的基础上，再进行信息抽取和信息查询，系统就可以提供更为精确的检索结果，提供更有价值的信息。

参考文献：

1 吴建中. DC元数据. 上海：上海科学技术文献出版社，2000

2 张 璞，庄成三. XML查询语言技术与实例分析. 计算机应用研究，2000(5)：109－111

3 刘 芳等. 半结构化、层次数据的模式发现. 小型微型计算机系统，2001，22(1)：84－88

（下转第102页）

到较为满意的智力和信息服务，又能使服务主体得到应有的收益。

鉴于咨询服务的高效益，用户愿以较大投入投向咨询业。像 AT&T、P&G、TENNECO 等著名的大型跨国公司，一年用于咨询业务的支出达数百万美元至数千万美元不等。印度在制定工业发展计划、保加利亚在发展能源工业、中国在建设三峡水库时，都曾由政府出面聘用国际上知名的大型咨询公司提供服务。与此相对应，咨询公司亦获得较高的经济收益。如：中国对外经济贸易咨询公司在 20 世纪 90 年代初的年收入达 650 万元；美国十大咨询公司的年收入达数千万美元至数十亿美元，年营业额平均增长率在 30% 以上，有的年份甚至高达 128%；斯坦福研究所在 20 世纪 70 年代、80 年代、90 年代的年收入分别为 0.7 亿美元、0.85 亿美元、1.5 亿美元。鉴于现代咨询业的高收益与高增长和它为客户带来的巨大效益，有人将其称之为"现代点金术"。

我国正处于经济社会发展的转型时期，对有效咨询服务的需求特别强烈。但是，我国咨询业发展面临着社会和体制条件约束以及行业自身方面的问题，如社会认同程度低、行业的市场化水平差、体制尚未理顺等，这就要求政府、咨询机构、研究与教学人员共同努力，顺应世界咨询业发展的潮流，尽快改变我国咨询业发展严重落后的不利局面，推动我国咨询业迅速成长壮大，以服务于社会的发展与进步。

参考文献：

1 上海市科学技术委员会. 上海市重大软科学成果选编(1983－1987). 上海：上海科学技术出版社，1989：77－81
2 龙永图. 关于经济全球化问题. 光明日报，1998－10－30
3 吴新年. 现代咨询产业及其发展动向研究. 情报杂志，1993(2)：5－9
4 李 楠. "点子公司"热度不减，咨询行业景气空前. 经济参考报，1994－07－28
5 加入 WTO，咨询业怎么办?. 人民日报(海外版)，2000－10－29

[作者简介] 江三宝，男，1961 年生，副教授，发表论文 20 余篇，出版专著 3 部。
姚云鸿，男，1965 年生，副教授，发表论文 10 余篇，出版专著 1 部。
王君南，男，1963 年生，副教授，发表论文 20 余篇，出版专著 4 部。
郭砚常，男，1963 年生，副教授，发表论文 10 余篇，出版专著 1 部。

(上接第 72 页)

4 王 静，孟小峰. 半结构化数据的模式研究综述. 计算机科学，2001，28(2)：6－10
5 魏定国. 半结构化数据库中的交互式查询和搜索. 计算机工程与应用，1998(9)：6－8
6 许学标等. 半结构化数据模型及查询语言. 计算机研究与发展，1998，35(10)：896－901
7 王蒙智等. 半结构化数据视图的增量维护. 计算机研究与发展，2001，38(2)：163－169
8 黄豫清等. 构造 Web 文档中半结构化信息的技术. 计算机辅助设计与图形学学报，2000，12(3)：230－234
9 陈 滢等. 基于标记图的 Web 数据模型. 计算机学报，1999，22(3)：306－312
10 陈恩红. 网际网上半结构化数据抽取与知识发现方法及其实现. 计算机科学，1999，26(10)：49－52
11 李 曦. 一种办公自动化中决策支持系统模型的研究. 中南工学院学报，1997，11(1)：52－57
12 李庆华，刘 昊. 用待确定的上下文无关文法分析半结构化数据. 华中理工大学学报，1999，27(5)：60－62
13 孟小峰. Web 数据管理研究综述. 计算机研究与发展，2001，38(4)：385－395
14 徐贵红. Web 的半结构化数据模型和查询模型. 蒙古大学学报(自然科学版)，1999，30(3)：299－303
15 刘 芳，胡和平. 半结构化数据的模式发现. 微型电脑应用，2000，16(2)：13－15

[作者简介] 黄 奇，男，1961 年生，副教授，硕士生导师，发表论文 20 余篇，出版专著、译著 7 部。
李 伟，男，1977 年生，硕士研究生。
接晓莉，女，1978 年生，硕士研究生。