

● 严亚兰

基于 Web 环境的半结构化数据模型研究 *

摘 要 半结构化数据的特点有:隐含的模式信息;不规则的结构;模式信息量大;模式变化快;先有数据,后有模式;模式是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;没有严格的类型约束;结构不完全。参考文献 13。

关键词 半结构化数据 XML 数据模型 数据库 检索机制

分类号 G250.74

ABSTRACT In this paper, the author summarizes the following characteristics of semi-structured data: irregular structure, mass quantity of schema information, rapid change of schema, data existing before schemas, description instead of constraint of data structures, etc. 13 refs.

KEY WORDS Semi-structured data. XML. Data modeling. Database. Search mechanism.

CLASS NUMBER G250.74

1 引言

WWW 的出现极大地丰富了因特网的内涵,促进了网络信息管理的实用化进程。网上内容以 HTML 格式表现和传输,网上现存的浏览器、Web 服务器和数据库服务器 3 层体系结构体现了人们设法利用数据库技术管理网上信息的努力,并且对以数据库方式存贮的信息进行管理具有很好的效果,但也有一些局限。(1)增大了网上传输的压力。因为浏览器只起到浏览信息的作用,无法对信息进行组织和管理;信息在数据库服务器中被组织检索,然后通过某种脚本语言(如 Vbscript 或 Javascript)将数据库格式的信息转变成 HTML 格式,再发送给浏览器。因此即便是用户非常简单的检索任务,也必须先传输给 Web 服务器,Web 服务器再通过数据库服务器将检索得到的结果传输给浏览器用户,这必然使网上通信能力降低。(2)对 HTML 格式信息管理力度不够。网上存在着巨大的 HTML 格式的信息源,对这部分信息的检索现在主要依赖搜索引擎,但搜索引擎的索引库大都是基于页面关键词的,查准率和查全率都不能令用户满意。(3)HTML 格式页面全文检索带来的问题。HTML 格式缺乏逻辑结构,例如要从某个网页中找到所有武汉市学生的信息,就需要编一个脚本程序,并逐字寻找“武汉”这个字符串,但检索结果并不都是居住在武汉市学生的信息,因为不住在武汉,但在某方面涉及到“武汉”的学生的信息也会被检索出来。

Web 数据管理中的数据组织是研究 Web 信息的特点,找出适合 Web 信息的合理组织模式。目前,基于 HTML 的文档型模式管理网上信息,暴露出越来越

越多的缺陷,人们迫切需要利用数据库对网上信息进行管理。如果网上传输的信息本身具有一定的逻辑结构,那么用户不必依赖于数据库服务器的管理,这时浏览器本身除了浏览功能外,同时可以对具有一定逻辑结构的信息直接进行组织和检索。但目前网上传输的是一种标记文本格式 HTML,如何使信息既能有效地在网上传输又具有一定的逻辑结构,这正是人类在 WWW 诞生和发展中致力解决的研究课题。目前的研究成果主要体现为半结构化数据模型的研究,而半结构化数据模型在网上的应用当首推 XML (eXtensible Markup Language) 语言。在 XML 数据中,搜索可以简单高效地进行,搜索引擎没必要再去遍历整个 XML 文件,它只须去找一找相关标记下的内容就够了。在上面提到的那个例子中,要想找“武汉”,只要看看<城市>这个标记下的字符串数据是否匹配就行了。半结构化数据模型研究及 XML 语言的开发,包括半结构化数据模型、数据查询、数据存储和交换等方面,其中数据查询是指能根据更丰富的语义信息在有效数据模型下找出更准确的信息。

2 半结构化数据模型研究的背景

在半结构化数据研究中,存在这样一个有趣的现象:一方面,在数据库研究领域,为了更有效地交换和传输数据,学者们试图用松散描述或不规则结构来表示数据,允许缺少模式(schema-less)的描述格式。在这种模式中,数据比起通常的数据库模式中的数据受到较少的约束。另一方面,在文档研究领域,为了有利于数据的组织和管理,学者们试图在文档中添加更多的结构,XML 标准的推出便是这种努力的结果。半结构化数据模型研究体现出这两种趋

*本文系教育部人文社会科学重点研究基地重大项目“信息可视化与知识检索”研究成果之一。

势的交叉融合。半结构化数据研究最早源自于数据库领域的学者,OEM(Object Exchange Model)模型和基于 OEM 的 Lorel 查询语言是半结构化研究中具有代表性的奠基成果。OEM 模型最初由 Papakonstantinou Y. 等人于 1995 年在其论文 *Object Exchange Across Heterogeneous Information Sources* 中提出,是用来表示半结构化数据的一个非常有用的、灵活的简单模型。它被作为美国斯坦福大学数据库小组(Stanford University Databases Group)各项研究项目的基础数据模型,Loirel 查询语言是该组 Lore 研究项目小组在 OEM 模型上提出的半结构化数据的查询语言。Lore 系统可从网上免费下载。从 1997 年推出 version1.1 开始,经不断升级,到 2000 年出台最新版 version5.0。Lore 系统有单用户版和多用户版,可在 Linux OS 或 Sun OS 上运行。

90 年代初,数据库领域面临的一个重大挑战是如何处理异构数据源的集成问题,因为企业越来越倾向于同时利用多种相互冲突的数据模型和模式来表达数据,而用户又希望通过一个集成的、一致的界面来存取所有的数据。这方面的研究已取得了很大成果,如 SQL3。但是,当用户希望通过一集成界面存取不仅仅存储在标准 SQL 数据库内的数据(信息),同时也存储在如对象仓库、知识库、文件系统和文档检索系统内的数据的时候,数据集成问题便显得更加具有挑战性。这种完全异构数据的集成与传统的数据集成虽然存在某些相似之处,但也表现出以下明显的区别:

(1)许多数据源包括的数据是非结构化或半结构化的,不存在规则的模式来描述数据,如可能存在自由文本(虽然在文本内也存在一些结构,但“字段”会以无法预料的方式发生变化)。

(2)环境是动态的,数据源的数量、内容和内容的含义会不断变化。

(3)信息存取和信息集成交互进行。在传统集成环境下,存在两个阶段:不同数据模式合并的集成阶段和数据的存取阶段。在新的环境下,在看到数据之前并不知道如何合并数据模式,而且模式集成方式会随着新的数据的出现而发生变化。

为了交换数据对象,任何一个信息系统都需要有一个一致的数据模式来表达对象,确定每一个对象的语义,确定对象在网上如何传输等。OEM 正是这样的一个对象交换模型。

3 以 OEM 为代表的半结构化数据模型基本理论

3.1 半结构化数据的特点

半结构化数据既不是天然数据(raw data,如图像

或声音),也不是常规数据库系统中严格结构化的关系数据库中的表或面向对象数据库中的有序图。它具有一定的结构,但其结构与数据混在一起,没有显式的模式定义。半结构化数据具有以下特点:隐含的模式信息;不规则的结构;模式信息量大;模式变化快;先有数据,后有模式;模式是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;没有严格的类型约束;结构不完全。

3.2 半结构化数据模式的抽取

数据库系统具有良好的图形界面和有效的存取方式,这主要得益于数据规则性的存在。利用已有的结构,用户和开发者可了解数据集并方便地存取数据;通过现有结构的优点(如利用索引),可使运行得到改善。虽然人们不得不处理网上没有明显结构且不规则的数据,但更愿意处理规则的、结构化的数据。这正是为什么要进行半结构化数据隐式模式抽取的原因。模式抽取研究中具有代表性的是基于 Datalog 检索语言的近似数据类型抽取。

3.3 半结构化数据模式的查询

为适应半结构化数据模式庞大且不完全被用户了解的特点,查询语言支持通配路径和正则路径的查询表达式。半结构化数据的查询本质上是对有向图的搜索,它具有以下特点:由于数据和模式信息统一存放(自描述性),与关系数据库查询相比,其优化处理较难;由于数据结构不规则或不完整,须支持数据类型自动转换(coercion);有向图中可能存在圈,因此在处理通配路径时,应避免查询进入无限循环,如在 Lorel 检索表达式 `Guide. restaurant. # @P. comp %, name` 中,先不考虑 `@P`,该表达式是指这样的路径:从 `Guide` 出发,经过 `restaurant` 边,经过有限的未特别指明的路径(符号 `#`),再经过具有开始字符为“`comp`”(`comp %`)的边,最后经过 `name` 边。`P` 是路径变量,通过 `@P` 与每一个与“`#`”相匹配的数据路径相绑定。变量 `P` 可以防止产生重复路径或进入无限循环。

4 基于 Web 环境的半结构化数据模型

以 XML 为基础的新一代 WWW 环境是直接面对 Web 数据的,不仅可以很好地兼容原有的 Web 应用,而且可以更好地实现 Web 中的信息共享与交换。XML 可看做为一种半结构化的数据模型,能够有效地对信息进行组织,并实施精确查询。

4.1 Web 环境的数据特点

(1)从广义数据库研究的角度看,Web 上的信息可看做一个更大、更复杂的数据库。Web 上的每一个站点就是一个数据源,不同的数据源都是异构的,因而每一站点之间的信息和组织都不一样,这就构

成了一个巨大的异构数据库环境。要想对这些数据进行组织,首先,必须研究站点之间异构数据的集成问题,只有将这些站点的数据都集成起来,提供给用户一个统一的视图,才有可能从巨大的数据资源中获取所需的東西;其次,必须解决 Web 上的数据查询问题,如果不能有效地得到所需的数据,就谈不上对这些数据进行分析、挖掘和预测。

(2) Web 上的数据与传统的数据库中的数据不同。传统的数据库都有一定的数据模型,可以根据模型来具体描述特定的数据。而 Web 上的数据非常复杂,没有显式的模式描述,每一站点的数据都各自独立设计,并且数据本身具有自述性和动态可变性。因而,Web 上的数据具有一定的结构性,但又是一种非完全结构化的数据,即半结构化数据。半结构化是 Web 上数据的最大特点。

4.2 XML 与半结构化数据

XML 数据与半结构化数据之间的对应是明显的,可以说 XML 是半结构化数据的一个特例。许多半结构化数据模型的研究可以很容易地应用到 XML 数据上。例如 OEM 模型可以作一些变动,用来表示 XML 数据。斯坦福大学数据库组将其开发的原型系统 Lore 在 Version 5.0 版本中移植至完全基于 XML 的数据模型上,并相应地扩展了 Lorel 查询语言。为了更有效地进行 XML 数据的处理,学者们提出了许多关于其模式描述的方案,最主要的是文档类型定义 DTD(Document Type Definition)。与半结构化数据的模式相比,DTD 作为模式的优点是它的正则语法支持定义半结构化数据,一个 DTD 可通过正则语法定义元素的嵌套结构,如下面 DTD 的例子。

```
例 1 <! ELEMENT part (name, brand, part)>
      <! ELEMENT name (# PCDATA)>
      <! ELEMENT brand (# PCDATA)>
```

其中每一个 part 元素可以包含任意深度的嵌套部分。另外,DTD 的其他语法也说明了它定义的是半结构化的模式,如<! ELEMENT e (a, b?, c+)>,表明元素 e 有一个子元素 a, b 子元素可选, c 子元素要求出现或连续出现多次。

4.2.1 XML 的查询语言

XML 查询语言一般通过两种途径来进行研究:第一,以 XML 数据形式为基础,将其语法进行适当的变形,成为一种便于使用的查询语言,如 XQL 语言。第二,以 SQL 或 OQL 为基础,添加必要的机制,使其能够表达一组有用的查询,如 XML-QL 语言。

XQL 是由 Microsoft、Texcel 等公司提出的一种 XML 查询语言,它是对 XSL (eXtensible StyleSheet Language) 的一种自然的扩充。XSL 模式语言是说明性的,而非过程性的。但 XSL 有许多不足之处,如不支持表达式,不能在节点间进行等值连接,对多个分散

的 XSL 文档没有一个形式化的查询机制,不支持聚类操作等。XQL 则在 XSL 基础上提供了筛选操作、布尔操作、对节点集进行索引,并为查询、定位等提供了单一的语法形式,因而在一定意义上可将 XQL 看做 XSL 的超集。XQL 设计组在设计之初便提出了以下 12 个设计目标:字符简洁易读;语法简单;以易于嵌入程序、脚本语言、XML 或 HTML 中的字符串表示;容易被解析;字符应与 URLs 自然匹配;应能描述 XML 文档中可能的任意路径和路径中任意的节点集;应能惟一标识 XML 文档中的每一节点;能返回任意数量的结果(包括 0);是说明性的,而非过程性的,如只指明查询什么,而不必指明如何查询,这一点很重要,因为检索优化器必须能自由地使用索引或其他结构来有效查询结果;应可起始于文档的任何层次位置而不必总是从根节点开始导航;检索结果应按文件排序,且没有重复节点;是用来从文档中检索信息,检索的信息可能是节点集、有关节点关系的信息或导出的值,语法中不指明输出格式,检索结果可能是一个节点、节点列表、XML 文档、一个数组或其他结构。XQL 不是以二进制而是逻辑结构返回结果。

XQL 主要针对以下几个问题而提出:

(1) 在单个 XML 文档中进行查询,如在 XML 浏览器和编辑器中对所处理的 XML 文档进行查询。另外,脚本语言也能利用这种性质给出一个非过程性的存取文档数据和结构的方法。

(2) 在 XML 文档集中进行查询。如在 XML 文档仓储 (Repository) 中进行查询。

(3) 能在 XML 文档间对节点进行定位。在 HTML 文档中,常常用 HyperLink 来定位其他文档,而在已有的 XML 中,链接形式更多样化,XQL 则主要是想通过给出已知位置的相对路径,或通过一绝对路径来在文档中定位任一节点。

XQL 的基本语法模仿了 URI (Unified Resource Identifier) 目录导航语法。通过 Context, XQL 可以指定 XQL 查询是以 XML 树结构中哪一层开始作为查询范围(例如,从树的根节点或儿子节点)。XQL 中以 / 表示根环境 (Root context), 以 ./ 表示当前环境 (Current context)。XQL 在 IE5.0 及以上版本中已得到部分应用。

XML-QL 是由 AT&T 实验室为首的数据库领域研究机构提出的 XML 查询语言。与 XQL 不一样,XML-QL 是面向 SQL 的,与 SQL 语言中的 SELECT 语句中的 WHERE 子句格式比较相像,并借鉴了一些半结构化数据查询语言的思想。在 XML-QL 中,将 XML 文档集看做数据库,以半结构化数据库模型为数据模型,将 DTD 文档看做数据库模式。

XML-QL 主要是针对以下几个方面而提出的:

(1) 能从大量 XML 文档中析取数据。

(2)能对 XML 数据进行转换。例如,将附有某一 DTD 文档的 XML 文档转换成为具有不同 DTD 文档的 XML 文档。

(3)将 XML 数据在多个 XML 数据源中进行集成。这对于 XML 数据仓储、数据挖掘等均很重要。

(4)客户端通过发送查询到 XML 数据源(可以是一个 Web 站点,也可以是多个 Web 站点)存取大量的数据。

4.2.2 XML 与 OEM 的比较

作为文本语言的 XML,它的基石是嵌套的标签对元素,每一个标签具有连续的零个或多个属性/值对和连续的零个或多个子元素,这些子元素本身可能也是带标签的元素,或者是“无标签”的文本数据部分。因为 XML 被定义为文本语言而不是数据模型,因而 XML 文档总是具有固有的顺序——这种顺序可能相关,也可能不相关,但不管怎样在一个文本表示中总是存在。一个结构良好的 XML 文档对标签、属性名或嵌套形式不加任何限制。如果一个 XML 文档通过 DTD 来限制其标签和结构,则只有满足 DTD 语法的 XML 文档被认为是有效的。由于 XML 不是一个数据模型,在实际应用中提出了 DOM (Document Object Model) 概念,它可将一个 XML 文档转变成数据结构形式以作为其他应用程序对 XML 调用的入口点。

XML 与 OEM 之间的区别主要表现在:

(1)XML 具有属性而 OEM 没有,XML 标签具有顺序而 OEM 没有。

(2)在对标签(tags/labels)的处理上:在 OEM 中,标签是作为入口点并标识与其他对象的关系——每一个 OEM 对象都必须具有一个标签(label);相反,XML DOM 强调每一个元素包含一个标签(tag)。

(3)XML DOM 至今不直接支持图结构,这无疑表明了 XML 文档的人为倾向。现在,XML 利用特殊的属性类型来编译图结构,每一个元素可具有一个惟一的 ID 类属性,其值提供了一个惟一标识符,该元素通过这个标识符被其他元素通过 IDREF 或 IDREFS 类属性引用。看下面 XML 的例子:

例 2 `<Person Id = "P1" Name = "Yan Yalan" Collaborator = "P2" >`

`<Person Id = "P2" Name = "Xiao min" Collaborator = "P1" >`

`<Publication Title = "The Information Mining on Internet" Author = "P1 P2" >`

假设属性 Id 是 ID 类型, Collaborator 是 IDREF 类型, Author 是 IDREFS 类型,上例中 Collaborator 和 Author 属性被看成是 Person 元素的标签引用,这类类似于 OEM 中的带标签的子对象。

5 结语

XML 自正式公布标准到现在已有 4 年多的历史,但由于 XML 的编辑、浏览工具的缺乏,致使 XML 的发展速度缓慢。另外,XML 的相关方案还在不断完善更新,尤其是 XML 的检索语言的开发还有待努力。应该说 W3C 组织为 XML 的开发和发展做出了卓越贡献,正如 W3C 站点首页所显示的座右铭“领导网络达到其最大潜力的使用”一样,正是 W3C 组织,使人类在半结构化领域研究的很多成果转变成网络环境下的应用变为可能。W3C 组织的 XML 查询工作组,号称他们的使命是提供灵活的查询以从 Web 上真实的和虚拟的文档中提取数据,提供 Web 世界和数据库世界之间所需要的交互,最终使 XML 文档能够像数据库一样对数据进行有效存取。

参考文献

- 1 <http://www-db.stanford.edu/lore/release/data.html>
- 2 Pajpakonstantinou Y., Garcia-Molina H., and Widom J.. Object exchange across heterogeneous information sources. <http://www-db.Stanford.edu/pub/papers/>
- 3 Abiteboul S. Querying semistructured data. <http://www-db.Stanford.edu/pub/papers/>
- 4 Nestorov S., Abiteboul S., Motwani R.. Extracting schema from semistructured data. <http://www-db.Stanford.edu/pub/papers/>
- 5 Buneman P., Davidson S., Fernandez M et al. Adding structure to unstructured data. <http://www-db.Stanford.edu/pub/papers/>
- 6 Abiteboul S., Quass D., McHugh J., Widom J., Wiener J.. The Lorel query language for semistructured data. <http://www-db.Stanford.edu/pub/papers/>
- 7 Abiteboul S., Widom J., Lahiri T.. A Unified Approach for Querying Structured Data and XML. <http://www-db.Stanford.edu/pub/papers/>
- 8 孟小峰. Web 数据管理研究综述. 计算机研究与发展, 2001(4)
- 9 McHugh J., Widom J.. Query optimization for XML. <http://www-db.Stanford.edu/pub/papers/>
- 10 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)
- 11 Goldman R., McHugh J., Widom J.. From Semistructured Data to XML: Migrating the Lore Data Model and Query Language. <http://www-db.Stanford.edu/pub/papers/>
- 12 <http://www.xml.net.cn>
- 13 <http://www.w3.org>

严亚兰 武汉大学信息管理学院博士生。通讯地址: 武汉市。邮编 430072。

(来稿时间: 2002-10-25)