一种查询 XML 异构数据源的新方法*)

陈金辉[1] 董 飚[2] 孙亚民[2]
(南京信息工程大学信息与控制学院 南京 210044)[1]
(南京理工大学计算机科学与技术学院 南京 210094)[2]

摘 要 由于异构数据源存在结构差异和结构不兼容等问题,在其上进行查询是一个挑战。本文根据 XML 树的特点,对其进行了外延,设计了一种新的 XML 树的查询方法。通过样式图获得 XML 树的结点间的语义关系,查询条件可以表示为 XML 样式图模式,查询不被限定于特定的 XML 树,给出了基于样式图模式的查询算法。用例说明了该方法如何应用于异构数据源的查询。

关键词 XML 树,样式图,样式图模式

A Novel Approach to Query Heterogenous Data Sources Based on XML

CHEN Jin-hui[1] DONG Biao[2] SUN Ya-min[2]
(School of Information & Control, Nanjing University of Information Science & Technology, Nanjing 210044, China)[1]
(School of Computer Science & Technology, Nanjing University of Science & Technology, Nanjing 210094, China)[2]

Abstract Querying heterogenous sources is a challenging issue due to structural differences and structural inconsistencies. Based on the characteristics of XML and several extensions on the XML tree, a novel querying approach is proposed for XML trees. A semantic relationship between nodes in XML trees is captured by the concept of a schema graph. Queries are specified on the schema graph patterns of the schema graph. A query is not restricted by the structure of a specific XML tree. A technique for evaluating queries is presented. Experimental results show how the approach can be used to query multiple XML trees in the presence of structural differences and inconsistencies.

Keywords XML tree, Schema graph, Schema graph pattern

1 引言

异构数据集成系统为访问异构数据源提供统一接口,使得查询者可以执行统一查询而不必考虑异构数据源之间的差异,异构数据源上的查询成为人们研究的热点。目前,XML 已成为 Internet 中数据表示和交换的事实标准,将 XML 技术引入到数据集成领域很有意义。但是,XML 异构数据源由于存在命名差异、结构差异和结构不兼容等问题,在其上进行查询是一个挑战性问题。

(1)XML 文档本质上可以看作标签在节点上的树。由于树形结构本身缺少语义信息,导致了命名差异问题。例如:笔记本在一个商品目录里属于膝上电脑,但在商品的另一个目录里又作为便携设备。

(2)在 XML 树形结构中组织同样的数据可能采用不同的方法,从而出现结构差异和结构不一致。例如:商品信息中有的包含销售商目录项,有的不包含,从而导致结构差异。又如:可以采用多种方式组织商品信息,如按厂家、按销售商、先按厂家再按销售商(或者先按销售商再按厂家)、不进行分类(厂家、销售商都是商品的孩子)等方式,从而导致结构不一致。

使用一些样式查询技术可解决第(1)种情况下的问题[1],本文研究第(2)种情况下的查询。查询 XML 具有结构差异和不一致问题的 XML 异构数据源有两种方法:一种是为初始的查询生成不同的版本。由于需要生成大量的查询,这种方法效率不高。另一种方法是先建立一个全局的结构,再在全局结构和本地结构之间建立映像规则。主要包括:Xyleme[2],Agora[3],文献[4-6]中分别提出了三种建立映像规则的方法。这些方法要求大量的手工操作,建立全局样式困难,并且映像规则对应用程序而言是"硬编码"的。

与本文相关的其它工作包括:Xclust[7]支持自动抽取本地数据源样式,文献[8,9]中研究了整合 DTD 的方法,文献[10-12]利用相关目录的语义信息提出了基于槽的层次化分类方法,这些工作都没有提供查询处理。另外,文献[13,14]中使用图样式来进行半结构化数据的查询,没有提出针对 XML 异构数据源的解决方案。

与上述方法相比,本文根据 XML 的特点对其进行了外延,提出了一个新的查询 XML 异构数据源的方法。主要特点有:

(1)查询条件定义在样式图上,不需转换到特定的 XML 数据源上,利用类型之间先序关系生成路径表达式;

(2)通过在异构数据源上建立关于查询的可映射集,实现了不需要通过"硬编码"就可对 XML 异构数据源查询;

(3)分析了查询的不可满足条件。

2 基本概念和定义

定义 1 一个 XML 数据树 T 为一个六元组:$T=(N,E,$

*)本研究得到南京信息工程大学科研基金资助项目(Y640)资助。**董 飚** 博士生,主要研究方向为中间件技术;**陈金辉** 副教授,主要研究方向为系统分析与集成;**孙亚民** 教授,博士生导师,主要研究方向为网络工程。

$D,V,Fnid,Fntype$),其中:

(1)N 是结点集合,$E\subseteq N\times N$ 是边的集合,D 是结点类型的集合,V 是结点标识的集合。

(2)$Fnid:N\rightarrow V$ 是结点集合到标识集合的映射,用于对每一个结点赋一个标识。兄弟结点的标识互不相同。

(3)$Fntype:N\rightarrow D$ 是结点集合到类型集合的映射,用于对每一个结点赋一个类型。各结点的类型是唯一的。

(4)$\exists!\ w\in N, Fntype(w)=R\wedge Fnid(w)=r$,该结点为根结点,它描述整个文档的全局信息,例如文档类型、发生时间等。

(5)从根结点开始的路径中的所有结点类型互不相同。

XML 按照树型结构来表达事实,本文在 XML 数据规范的基础上进行了一些外延,以提高查询 XML 异构数据源的能力。

定义 2 在 T 上定义的样式图是一个四元组:$GS_T=(N^{GS},E^{GS},V^{GS},Find^{GS})$,其中:

(1)N^{GS} 是结点的集合,并且 T 的根结点 $R\in N^{GS}$,称为主结点;

(2)$E^{GS}\subseteq N^{GS}\times N^{GS}$ 是结点之间的有向边的集合,并且有下式成立:

$\forall e=\langle N_i,N_j\rangle\in E^{GS}\Leftrightarrow\exists n'_i,n'_j\in N\wedge Fntype(n'_i)=N_i\wedge Fntype(n'_j)=N_j\wedge N_i\in D\wedge N_j\in D\wedge n_i=Parent(T,n_j)$

函数 $Parent(T,x)$求出 T 中结点 x 的父结点;

(3)$V^{GS}\subseteq D$ 是结点标识的集合,并有 $\forall x\in V^{GS}\Leftrightarrow\exists y\in V\wedge Fntype(y)=x$;

(4)$Find^{GS}:N^{GS}\rightarrow V^{GS}$ 是结点集合到标识集合的映射,用于对每一个结点赋一个标识。

样式图规定了 XML 数据树的形状以及对结点和边的约束,可用有向图来表示。样式图中的结点和边与数据树中的结点和边具有对应关系。样式图中存在一个主结点,有向边表示元素的嵌套包含关系。

定义 3 在 GS_T 上定义的样式图模式 PGS_T 是一个两元组:$PGS_T=(F^{PGS},PRE^{PGS})$,其中:

(1)$F^{PGS}=\{filter_i(N_i^{GS})\mid 0\leqslant i\leqslant|N^{GS}|\}$是过滤函数的集合,其中过滤函数

$filter_i(N_i^{GS})=\{\langle V_i^{GS}=B_i\rangle\mid Fntype(N_i^{GS})=V_i^{GS}\wedge B_i\subset V\wedge B_i=\text{“?”}\}$

$filters_i$ 表示 N_i^{GS} 应满足的约束条件,B_i 是 V 的子集,或者为“?”,当为“?”时,其作用相当于通配符。当 $filter_i\neq\varnothing$ 时,称 GS_T 上对应结点 N_i^{GS} 被标注。F^{PGS} 可为空集;

(2)PRE^{PGS} 是样式图中具有先序关系的结点组成的对偶的集合,其中每一项$\langle N_i^{GS},N_j^{GS}\rangle$表示样式图中从主顶点开始的包括 N_i^{GS} 和 N_j^{GS} 的路径中,N_i^{GS} 离主顶点距离更近。当 N_i^{GS} 和 N_j^{GS} 具有父子关系,记为$\langle N_i^{GS},N_j^{GS}\rangle\in PREP^{PGS}$;当 N_i^{GS} 和 N_j^{GS} 具有祖孙关系,记为$\langle N_i^{GS},N_j^{GS}\rangle\in PRED^{PGS}$。$PRE^{PGS}$ 可为空集。

样式图模式与 XML 异构数据源的查询语言对应,样式图模式规定了对结点的选择、约束条件以及结点间的先序关系。

定义 4 样式图模式 PGS_T 的查询结果是 T 的一棵子树 τ,满足:

(1)τ 和 T 有同样的根结点,τ 中的叶结点是 T 的叶结点;

(2)对于 GS_T 中被标注的结点 N_i^{GS},则在 τ 中从根到树叶的路径上的有一对应的结点 N_j,且 $N_i^{GS}=Fntype(N_j)\wedge filter_i(N_i^{GS})$;

(3)若 PGS_T 存在,则$\langle N_i^{GS},N_j^{GS}\rangle\in PRE^{PGS}$,则在 τ 中从根结点到树叶的每一条路径上与其对应的结点之间也存在先序关系;

(4)τ 是包含最大结点数的子树,也称 τ 为 T 的最大子树。

条件(2)表明 τ 中对应的结点 N_j 应满足 PGS_T 的约束条件,在 PGS_T 中,若顶点 N_i^{GS} 的约束条件为 $V_i^{GS}=\text{“?”}$,表示在 τ 的 N_j 结点处,V_i^{GS} 的所有元素满足条件。τ 是 T 中从树根到树叶的满足条件的路径集合,该集合被聚合到一起构成 T 的一棵子树。

定义 5 当且仅当 τ 为空时,称 PGS_T 不可满足;否则,称 PGS_T 可满足。

定理 1 如果下列条件之一成立,则 PGS_T 是不可满足的。

(1)PGS_T 中的具有先序关系的顶点构成一个有向环;

(2)PGS_T 存在先序关系$\langle N_i^{GS},N_{j1}^{GS}\rangle$且$\langle N_i^{GS},N_{j2}^{GS}\rangle$,或$\langle N_{j1}^{GS},N_i^{GS}\rangle$且$\langle N_{j2}^{GS},N_i^{GS}\rangle$,则 $N_{j1}^{GS}\neq N_{j2}^{GS}$;

(3)存在$\langle N_i^{GS},N_j^{GS}\rangle\in PREP^{PGS}$,但$\langle N_i^{GS},N_j^{GS}\rangle\notin E^{GS}$;

(4)$\langle N_i^{PGS},N_j^{PGS}\rangle\in PRED^{PGS}$,但在 GS_T 中不存在从 N_i^{GS} 到 N_j^{GS} 的路径;

(5) 存在一个被标注的结点 N_i^{GS},该结点不在 GS_T 中从根结点开始的任何一条路径上。证明从略。

定义 6 GS_T 上关于 PGS_T 的查询路径集 $Path(PGS_T)$ 是 GS_T 上从根结点开始的路径的集合,其中任意一条路径 $path_i(PGS_T)$满足:

(1)对于任意被标注的结点都在该路径上,并且该路径的终结点为一标注结点;

(2)如果存在$\langle N_i^{GS},N_j^{GS}\rangle\in PREP^{PGS}$($\langle N_i^{GS},N_j^{GS}\rangle\in PRED^{PGS}$),则在该路径中 N_i^{GS} 是 N_j^{GS} 的父结点(祖先结点)。

定理 2 PGS_T 是不可满足的,当且仅当 $Path(PGS_T)=\varnothing$。证明从略。

定义 7 设 $R,N_1^{GS},\cdots,N_k^{GS}$ 是 PGS_T 的一条查询路径,其路径表达式为 $ROOT/\Omega_1/\cdots/\Omega_k$。当标注结点 N_i^{GS} 取值为 $\{V_1\mid\cdots\mid V_m\}$时,$\Omega_i=(V_1\mid\cdots\mid V_m)$;当在 GS_T 上 N_i^{GS} 未被标注或者标注结点 N_i^{GS} 的约束条件 $V_i^{GS}=\text{“?”}$时,$\Omega_i=*V_i^{GS}$。

定义 8 设 Φ 是具有相同根结点的一组 XML 数据树的集合(或者从根结点到树叶路径的集合),$\bigcup_{p\in\Phi}p$ 是包含 Φ 中的元素作为子树的结点最少的一棵 XML 数据树。

定理 3 $\bigcup_{p\in\Phi}p$ 是唯一的。证明从略。

定义 9 设 e 是 PGS_T 上的一个路径表达式,Pa 是 T 上从根到树叶,并且满足 e 的路径集合,T 上的一条查询路径表达式的结果 $Res(e,T)=\bigcup_{p\in Pa}p$。

查询路径表达式的结果 $Res(e,T)$不同于相应的 $XPath$ 路径表达式的结果,因为一个查询路径表达式的结果是一棵 XML 数据树,而相应的 $XPath$ 路径表达式的结果是结点的集合。

定理 4 τ 能通过合并 T 上相应的查询路径表达式的结果来计算。设 $PATH(PGS_T)=\{e_i\mid 1\leqslant i\leqslant n\}$,则 $\tau=\bigcup_{i\in[1,n]}res(e_i,T)$。证明从略。

算法 计算 τ。首先检测 PGS_T 的可满足性,如果可满足,再分三步进行:

(1) 计算查询路径 $PATH(PGS_T)$;

(2) 根据 $PATH(PGS_T)$,为每一条查询路径生成一个路径表达式;

(3) 在 T 上计算路径表达式的值,并且组合成 τ。算法的证明从略。

定义 10 设异构数据源上的一组 XML 数据树 $HT_i=(N_i,E_i,D,V,Fnid_i,Fntype_i)(1\leqslant i\leqslant n)$,对应的样式图记为 $GSHT_i$,则全局样式图 $GS_{\Sigma HT}=GSHT_1\cup\cdots\cup GSHT_n$。

定义 11 $GS_{\Sigma HT}$ 上的一个查询 $PGS_{\Sigma HT}$ 能被映射到 $GSHT_i(1\leqslant i\leqslant n)$ 上,当且仅当不存在出现在 $GS_{\Sigma HT}$ 上但不出现在 $GSHT_i$ 上的结点。满足 $PGS_{\Sigma HT}$ 映射条件的 $GSHT_i$ 组成的集合称为 $GS_{\Sigma HT}$ 上的可映射集,记为 $F(GS_{\Sigma HT})$;否则,称 $PGS_{\Sigma HT}$ 不能被映射到 $GSHT_i$ 上。

定理 4 在全局样式图 $GS_{\Sigma HT}$ 上的一个查询 $PGS_{\Sigma HT}$ 能被分解为 $F(GS_{\Sigma HT})$ 上的各样式图的查询。如果查询 $PGS_{\Sigma HT}$ 未被映射到某个样式图 GS_{HTi} 上,则返回空树作为查询结果。证明从略。

定理 5 定义在全局样式图 $GS_{\Sigma HT}$ 上的查询 $PGS_{\Sigma HT}$ 不需要通过映像规则映像到各自的样式图 GS_{HTi} 上就可实现对 XML 异构数据源的查询。证明从略。

3 用例分析

本节用图示说明文中提出的 XML 异构数据源的查询方法。图 1 为 4 棵 XML 数据树,图中树叶结点 Δ 标识父结点,虚线矩形方框表示一组具有相同类型的结点。这 4 棵数据树存在结构差异和结构不一致现象。

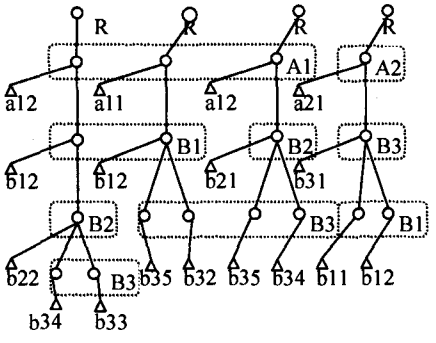

图 1 4 棵 XML 数据树

图 2 为对应的样式图,图 3 为对应的全局样式图及其上的模式,• 为标注结点,虚线箭头表示先序关系。图 3 中 A1,B1 和 B3 为标注结点,$F^{PGS}=\{B1=\{b12\},B3=\{b32,b33\}\}$,$PREP^{PGS}=\{\langle A1,B3\rangle\}$。

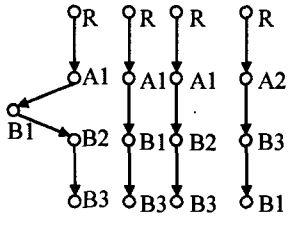

图 2 样式图

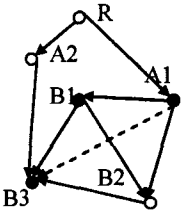

图 3 全局样式图模式

图 3 有查询路径 $/R/A1/B1/B2/B3/$,$R/A1/B1/B3$ 和 $/R/A1/B2/B3/B1$,对应的路径表达式(1)$/r/*A1/b12/*B2/(b32|b33)$,(2)$/r/*A1/b12/(b32|b33)$ 和(3)$/r/*A1/*B2/(b32|b33)/b12$。

图 4 为全局样式图模式在 4 棵数据树上的查询结果,路径表达式(1)和(2)的查询结果分别是图 3 中由根到叶的两条路径,路径表达式(3)的查询结果是一棵空树。

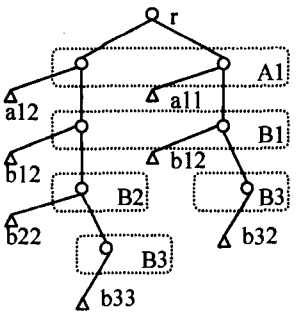

图 4 查询结果

结束语 针对现有 XML 异构数据源查询方法中的问题,提出了一种新的查询方法:根据 XML 数据树的类型信息抽取出样式图,在样式图上建立样式图模式,用户的查询条件与样式图模式对应,定义了查询结果的语义,给出了查询的算法。通过在异构数据源上建立关于查询的可映射集,实现了不需要通过"硬编码"就可对 XML 异构数据源查询。分析了样式图模式的不可满足条件。用例分析表明了该方法的有效性。

参 考 文 献

[1] Rahm E, Bernstein P A. A Survey of Approaches to Automatic Schema Matching. VLDB Journal, 2001, 10(4): 334-350

[2] Cluet S, Veltri P, Vodislav D. Views in a Large Scale XML Repository // Proc. of the VLDB'01 Conference. Rome, Italy, 2001: 271-280

[3] Manolescu L, Florescu D, Kossmann D. Answering XML Queries over Heterogeneous Data Sources // Proc. of the VLDB'01 Conference. Rome, Italy, 2001: 241-250

[4] Amann B, Beeri C, Fundulaki L, et al. Ontology-based Integration of XML Web Resources // Proc. of the ICSW'02 Conference. Sardinia, Italy, 2002: 117-131

[5] Christophides V, Cluet S, Simeon J. On Wrapping Query Languages and Efficient XML Integration // Proc. of the ACM SIGMOD'00 Conference. USA, 2000: 141-152

[6] Marron P J, Lausen G, Weber M. Catalog Integration Made Easy // Proc. of the ICDE'03 Conference. Bangalore, India (poster), 2003: 677- 679

[7] Lee M L, Yang L H, Hsu W, et al. Xclust: Clustering XML Schemas for Effective Integration // Proc. of the CIKM'02 Conference. Virginia, USA, 2002: 292-299

[8] Behrens R. A Grammar-based Model for XML Schema Integration // Proc. of the BNCOD'00 Conference. Exeter, UK, 2000: 172-190

[9] Garofalakis M, Gionis A, Rastogi R, et al. XTRACT: A System for Extracting Document Type Descriptors from XML Documents // Proc. of the ACM SIGMOD'00 Conference. Dallas, Texas, USA, 2000: 165-176

[10] Van Dijk P. eXchangeable Faceted Metadata Language-XFML Core, (2003). http://www.xfml.org/spec/1.0.html. (accessed August 2007)

[11] XML Topic Maps (XTM 1.0). 2001. http://www.topicmaps.org. (accessed August 2007)

[12] Tzitzikas Y, Spyratos N, Constantopoulos P, et al. Extended Faceted Taxonomies for Web Catalogs // Proc. of the WISE'02 Conference. Singapore, 2002: 192-201

[13] Buneman P, Davidson S B, Fernandez M F, et al. Adding Structure to Unstructured Data // Proc. of the ICDT'97 Conference. Delphi, Greece, 1997: 336-350

[14] Goldman R, Widom J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases // Proc. of the VLDB'97 Conference. Athens, Greece, 1997: 436-445