文章编号： 1000-5013(2012)05-0509-04

基于刻面的数据空间数据源管理子系统

王江海，武林仙，吴扬扬

(华侨大学 计算机科学与技术学院，福建 厦门 361021)

摘要： 提出一种基于刻面描述的数据空间数据源描述模型(FADSM)，实现数据空间对数据源“先有数据，后在模式”的即插即用的管理模式. 在数据空间原型系统架构下，以FADSM模型为基础构建一个数据空间数据源管理子系统. 原型系统实现对数据空间中异构异质数据源内部及外部属性的提取，以Pay-As-You-Go的管理模式实现统一管理，并提供对数据源添加、删除和浏览等基本管理功能.

关键词： 数据空间；刻面；数据源管理；异构异质数据

中图分类号： TP 311.13 **文献标志码：** A

信息技术与计算机网络的飞速发展，在实现数据共享的同时，也使用户不得不面对大量的不断快速增长的数据. 数据的海量、共享性及其多样性使得传统的关系数据库管理模式面临着严峻的挑战. 数据空间(dataspace)就是针对异构异质数据数据管理难的问题提出来的. 与关系型数据库区别的是，将数据加入到数据空间之前，无需像关系数据库事先为其定义关系模式，而直接将数据源加入数据空间，并以Pay-As-You-Go模式实现数据的管理[1-3]，使其更能适应未来各种异构异质数据的管理需求. iDM(imemex data model)[4]是通过资源视力来描述数据源，但基于iQL查询可能会很复杂；UDM(unified data model)[5]主要是关注桌面搜索的无法提供关系数据查询；Triple Model[6]是基于RDF的，提供了强大的查询能力，但不支持属性查询和不确定查询，普通用户使用比较困难；Probabilitstic Sematic Model[7]是基于概率的，能够处理不确定数据源，但其扩展性受到使用的集成方法的限制. 基于任务的数据空间模型[8]只是从用户任务方面考虑的，弱化了数据源内容；PAD和CKP模型[9]使用了本体的概念，但其本体本身的建构需要领域专家的参与；RSM(refined standard model)[10]将数据空间看作是若干个资源的空间的集合，各个资源空间中有相同属性的数据聚类，但却忽略了不同类数据间的内容间关联性；LGDM(layered graph data model)[11]也是基于图的模型，以对象的概念作为数据最小单元. 若干属性对数据源描述可以是对数据源的某方面特征的描述，而以上介绍的数据空间模型在描述数据源时多是将数据源看作简单的属性集合，忽略了属性间的关系. 为描述数据空间中的数据，本文提出一种基于刻面描述的数据空间模型(FADSM模型)，并在此模型上构建了一个数据源管理子系统.

1 数据空间数据源的描述模型

在软件构件库的分类模式中，刻面分类将对构件描述的关键词置于不同的语境，从而可以从多个视角来观察构件，以此来精确分类构件. 通常对数据源的描述是基于属性集合的，即通过属性名和属性值元组的集合来完成. 这种表达方式只是将数据源看做简单的属性集合，并没有进一步挖掘出属性间的关系. 文中对这些属性进行了进一步的抽象，提取属性之间的关系，将各个属性划入不同的刻面.

在基于刻面的概念下，通过数据源、刻面和属性来描述数据源. 数据源并非单独存在的，它同时与其它数据存在着各种各样的关联，如引用、具有相同的刻面等. 因此，在对数据源描述时不能仅描述数据源内部属性的关系，还需要引入一个关系集来描述各种不同数据源之间的关系. 即通过刻面、属性和关系

收稿日期： 2012-03-24

通信作者： 吴扬扬(1957-)，女，教授，主要从事数据库和数据挖掘的研究. E-mail：wuyy@hqu.edu.cn.

基金项目： 福建省科技计划重大项目(2011H6016，2011H0028)

来描述数据空间中的数据源.

定义如图1所示的数据空间的FADSM模型为Dsource=(ID,*FS*,*A-VS*). 其中:ID是数据源的标识符,表示数据的类别和存储位置,类似于URL的表示方式;*FS*是数据源的刻面集合;*A-VS*是刻面所包含的内容集合,包括了描述这个数据源的所有属性及关系等.

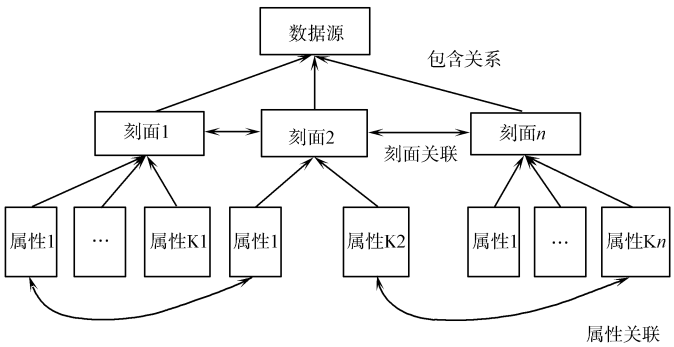

图1 FADSM模型示意图

Fig. 1 Diagram of FADSM model

在FADSM模型中,属性用来描述数据源对象的特性,如邮件用来描述文件的大小、位置、时间等,数据库的表、列等,网页的URL,Title等.对于用户对数据源的自定义属性,也用来作为对数据源的描述加以使用.刻面是指用户观察数据源的视角,如一张新闻网页,从文件的角度看,它有网页大小、网页存放位置、网页创建时间等属性;从内容的角度看,有新闻标题、新闻记者、发生时间等属性;而从网页的角度看,则有URL地址,Title,相关页等属性.本模型可以为用户提供从不同的角度浏览和查询数据.

在研制的数据源管理子系统中,对于常见的数据源,设置了两个初始的刻面集及其属性集,用户可根据需要增加或修改.初始的刻面包括Basic基础刻面和Content内容刻面. 表1是一些常见数据源的初始刻面集和属性集.

表1 常见数据源初始刻面及属性集

Tab. 1 Initial facets and attributes for common data sources

数据源	标识符(ID)	刻面集(*FS*)	属性二元组(*A-VS*)	
			属性名集(*AS*)	值集(*VS*)
文件	File-id	{Basic,Content}	{name,size,location,time,…},{title,author,abstrat,…},…	{file1,xx M,d:\.},{"about",zhao,…}
图片	Photo-id	{Basic,Content}	{name,size,…},{pixel,type,…},…	{photo1,xx M},{xx Px}
网页	Page-id	{Basic,Content }	{name,size,…},{url,title,…}	…
数据库	Database-id	{ Basic,Content}	{table,lumn,type,…}	…
邮件	Emailbox-id	{ Basic,Content}	{name,size,…}, {suject,sender,…}	…

数据源管理子系统对加入数据空间的数据源自动抽取其各个刻面的属性,并建立其多刻面地描述模式.用户不需要定义数据模式,就能对异构异质数据源的管理,实现数据源的浏览、查询和检索.

2 数据空间数据源管理子系统

基于上述FADSM模型,构建一个数据空间数据源管理子系统,如图2所示.系统通过对数据空间中异构、异质数据源内部及外部属性的自动提取,以Pay-As-You-Go的管理模式实现数据源的统一管理,并提供了对数据源添加、删除和浏览等基本管理功能,为将来数据空间索引及空间演化提供了基础.该系统主要由5个模块组成,包括显示模块、数据源管理模块、属性存储模

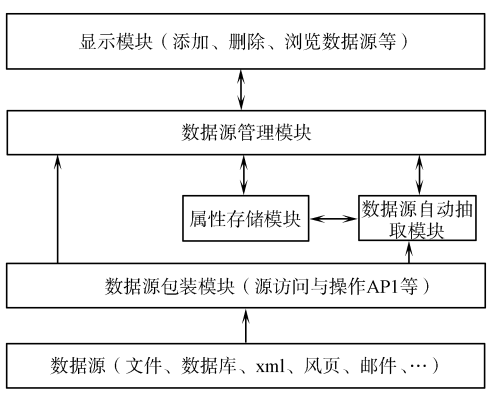

图2 数据空间数据源管理子系统架构图

Fig. 2 Architecture of data management subsystem for dataspace

块、数据源自动抽取模块和数据源包装模块.

1）数据源包装模块．主要定义了文件、数据库、邮件和 xml 等数据源的刻面描述信息的访问方法.模块向上提供对异构数据源的元数据信息及内容的访问接口，实现对数据源的统一访问，在后续的新数据源中只需要实现元数据访问接口就可以保证对新数据源的访问.

2）数据源管理模块．数据源管理模块提供数据空间中数据源管理的主要功能，完成包括添加数据源、删除数据源和浏览数据源 3 个主要业务逻辑.模块实现了数据源加入到数据空间、利用属性提取模块抽取数据源属性，以及向显示模块提供数据源刻面描述信息的查询及内部数据的浏览方法.

3）自动提取模块．包括对数据源外部信息和内容信息的刻面描述的自动抽取及数据源内容的索引.模块根据提供的数据源访问信息判断数据源类型，调用数据源包装模块提供的数据源访问 API，获取数据源的刻面描述信息并存储.外部属性的提取主要是针对数据源各种外部描述元数据的提取；对于内部内容信息的抽取，通过基于加权重规则统计、贝叶斯分类模型和支持向量机模型结合的机器学习方法，对标题、作者、关键字、主题和语言等数据信息进行提取，同时通过 Lucene 工具对数据源内容进行索引，方便查询.

4）存储模块．使用刻面描述模型对数据源以数据源、刻面、属性 3 个层次来描述并存储，并提供对这些信息的查询方法.这种存储方式与数据源本身的异构性无关，具有良好的扩展性能，对数据源信息的变更不影响存储的本身结构.同时，属性的存储的访问接口提供了对插入数据源属性到属性存储的访问方法，保证了属性自动提取模块的相对存储的独立性.

5）显示模块．提供用户将数据空间外部的数据源加入到管理子系统中、数据空间内部数据源移除据空间管理等操作的用户界面，并提供对数据源刻面描述信息的浏览（数据源的刻面名、刻面集合等）及数据源内容查看的界面.

数据空间数据源管理子系统的系统界面共分为 4 部分．最上层是系统的菜单栏和工具栏，提供数据空间原型系统的基本功能的入口，包括数据源添加和删除、数据源索引、数据空间配置等功能入口；左边树型结构区域是数据空间中数据源树型浏览区，提供数据空间中所有数据源的浏览入口；右边窗口上部是数据空间的查询入口，提供数据源空间的关键字查询；右下部是内容显示区，提供数据源内容、关键字查询、基于用户活动的查询结果等内容的显示.

在实验中，将 236 个普通文件（74.5 Mb）、262 封邮件（20.4 Mb）、8 个数据库（435.4 Mb）和 78 个 xml 数据文件（52.2 Mb）共 4 类异构异质数据源加入到数据空间中进行管理．用户通过菜单栏中的数据源菜单下的添加数据源菜单，进入数据源添加窗口；窗口提供了多种异构数据源的添加功能，用户只需要选择相应的数据源，并提供访问时所需要的连接信息；点击确定后，系统将在后台自动抽取数据源刻面信息，并对数据源内容进行索引.数据源加入数据空间后，用户浏览系统抽取的数据源的刻面描述信息，如图 3 所示.对于加入到数据空间数据源管理子系统的数据源，用户可以通过上面的查询框中，输入刻面信息的关键字来查询相关的数据源.

图 3　数据空间数据源刻面浏览

Fig. 3　Data resource facet browse for dataspace

3　结论

提出了一种基于刻面描述的数据空间数据源描述模型（FADSM），并在此基础上利用 Java 语言的优势构建了一个数据源管理子系统，实现了对数据空间中异构异质数据的统一管理.虽然系统未实现对空间中数据源变化的监控及空间的进化，但统一的管理方法及数据源的存储方法为将来数据空间索引及空间演化提供了基础.

实验结果表明:FADSM模型满足了数据空间对异构异质数据源的统一管理的要求.数据源管理子系统通过预先对数据源的基础刻面的抽取提供数据源的基本管理功能.下一步的工作,将是完成对数据源的监控和挖掘的数据源间关系,以实现数据空间的演化,为用户提供更强大的服务.

参考文献:

[1] FRANKLIN M,HALEVY A,MAIER D. From databases to dataspaces: A new abstraction for information management[J]. ACM SIGMOD Record,2005,34(4):27-33.

[2] HALEVY A,FRANKLIN M,MAIER D. Principles of dataspace systems[C]// 25th International Conference on Management of Data Principles of Database Systems. Chicago:ACM SIGMOD,2006:1-9.

[3] HALEVY A,FRANKLIN M,MAIER D. Dataspaces: A new abstraction for information management[C]// 25th International Conference on Management of Data Principles of Database Systems. Chicago:ACM SIGMOD,2006:1-2.

[4] DITTRICH J P,SALLES M A V. iDM:A unified and versatile data model for personal dataspace management[C]// Proceedings of the 32nd International Conference on Very Large Data Bases. Seoul:[s. n.],2006:367-378.

[5] PRADHAN S. Towards a novel desktop search technique[C]// Proceedings of 18th International Conference on Database and Expert Systems Applications. Regensburg:[s. n.],2007:192-201.

[6] ZHONG Ming,LIU Meng-chi,CHEN Qian. Modeling heterogeneous data in dataspace[C]// IEEE International Conference on Information Reuse and Integration. Las Vegas:[s. n.],2008:404-409.

[7] SARMA A D,DONG X L,HALEVY A Y. Data modeling in dataspace support platforms[J]. Conceptual Modeling: Foundations and Applications,2009,5600:122-138.

[8] 寇玉波,李玉坤,孟小峰,等.个人数据空间管理中的任务挖掘策略[J].计算机研究与发展,2009,46(增刊2):446-452.

[9] 董彦磊,申德荣,寇月,等.数据空间中数据组织模型以及关联关系发现模型的研究[J].计算机研究与发展,2009,46(增刊2):191-199.

[10] JIANG Xiao-rui,SUN Xiao-ping,ZHUGE Hai. A Resource space model for dataspace[C]// Sixth International Conference on Semantics, Knowledge and Grids. Washington D C:IEEE Computer Society,2010:33-41.

[11] YANG Dan,SHEN De-rong,NIE Tie-zheng,et al. Layered graph data model for data management of dataspace support platform[J]. Web-Age Information Management,2011,6897:353-365.

A Data sources Management Subsystem for Dataspace Based on Facets

WANG Jiang-hai, WU Lin-xian, WU Yang-yang

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: A facet-based attributes dataspace model (FADSM) is proposed in this article, which implements data-first management model. In the architecture of dataspace prototype, we design a subsystem for data sources management in dataspace based on FADSM. Our system achieves to extract the internal and external attributes of heterogeneous data in dataspace and manage data in Pay-As-You-Go style. It also implements the basic functions to add , delete and browse data sources in dataspace, which provides a basis for data indexing and evolution in dataspace.

Keywords: dataspace; facets; data source management; heterogeneous data

(**责任编辑**:陈志贤 **英文审校**:吴逢铁)