企业数据空间的数据组织方法研究

文必龙,焦圣杰,郭 娇
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:企业数据空间的主体是整个企业,面向多个部门、专业或项目,数据规模巨大、种类复杂多样,还需要管理大量已有严格数据模式的数据库,数据组织管理困难。然而现有数据空间的数据组织方法,大多是基于个人数据空间的,无法满足企业数据空间复杂的数据管理需求。为了统一描述多源异构数据、多维多角度地灵活组织数据和将传统的"先模式后数据"和数据空间的"先数据后模式"方式协调起来进行管理,提出了企业数据空间的数据组织方法:通过构建分层组织模型实现对数据多维多角度地灵活组织;利用属性图模型对企业数据空间中的各种数据资源进行统一描述和管理。基于该方法可以更好地描述企业中的各种数据资源,为企业提供灵活高效的数据组织和管理方式,进而更好地支持企业数据空间的数据模式演化,提高企业的数据管理效率,满足企业的数据组织管理需求。

关键词:企业数据空间;数据组织;数据资源目录;属性图;数据模型
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)12-0056-05
doi:10.3969/j.issn.1673-629X.2020.12.010

Research on Data Organization Method Based on Enterprise DataSpace

WEN Bi-long,JIAO Sheng-jie,GUO Jiao
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:The main body of the enterprise data space is the entire enterprise,facing multiple departments,professions,or projects. The scale of the data is huge,and the types are complex and diverse. It also needs to manage a large number of databases with strict data models,which makes data organization and management difficult. However,the existing data organization methods of data space are mostly based on personal data space,which cannot meet the complex data management requirements of enterprise data space. In order to uniformly describe multi-source heterogeneous data,organize data flexibly in multiple dimensions and multiple angles,and coordinate the traditional "first model first data" and data space "first data first mode" methods for management,a data organization for enterprise data space is proposed. A multi-dimensional and multi-angle flexible organization of data is achieved by constructing a hierarchical organization model; a unified description and management of various data resources in the enterprise data space is made by using the attribute graph model. Based on this method,various data resources in the enterprise can be better described,and flexible and efficient data organization and management methods can be provided for enterprises,so as to better support the data model evolution of the enterprise data space,improve the data management efficiency of the enterprise,and meet the needs of enterprise data organization and management.

Key words:enterprise dataspace;data organization;data resource catalog;property graph;data model

0 引 言

为了更好地应对数据的海量、异构、共享性和多样性给数据管理带来的挑战[1],Franklin 提出了数据空间(Data Space)的概念:一个数据空间由一系列相关的异构资源对象集和资源对象间的关联关系集组成,包含某个组织或个体相关的一切信息,这些信息可以以任意形式,在任意地方存储;在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严格的关系模式,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[2]。

有不少学者对数据空间技术开展了研究,其中最具有代表性的是个人数据空间技术,如瑞士苏黎世理

收稿日期:2020-02-28 **修回日期**:2020-06-30
基金项目:国家自然科学基金面上项目(41574117);国家重大专项(2016ZX05033-005-004);大庆市指导性科技计划项目(zd-2019-22)
作者简介:文必龙(1967-),男,博士,教授,硕导,研究方向为大数据、软件工程等;焦圣杰(1994-),男,硕士研究生,研究方向为软件工程、数据空间。

工学院开发的 iMeMex 系统[3]、美国华盛顿大学开发的 SEMEX 系统[4]以及中国人民大学开发的 OrientSpace[5]等个人数据空间系统原型。个人数据空间的许多技术同样适合企业数据空间,如个人数据空间模型与查询、数据关联、数据索引等,但与个人数据空间相比,企业数据空间管理更加复杂。

企业数据空间的主体是整个企业,而不是个人或某个部门。需要管理的数据来自各个部门、专业、项目或者业务,数据规模巨大、种类复杂多样,组织管理困难。而且,企业现有数据库具有专门的,甚至标准化的数据模式,与个人数据空间的数据模式的灵活性相反,这些数据库要求的是模式稳定,需要将传统的"先模式后数据"和数据空间的"先数据后模式"的方式协调起来进行管理。

针对灵活高效地组织企业数据空间中的数据资源的问题,该文结合企业数据管理的特点,提出了企业数据空间的数据组织方法:通过构建的分层组织模型实现对数据进行多维多角度地组织,利用属性图模型统一描述企业数据空间中的各种数据资源,实现了对企业数据灵活和高效的组织管理。

1　相关研究

目前,数据空间中数据组织方面的研究主要包括数据空间体系架构、数据空间数据模型表示方法、数据索引、数据关联关系挖掘等。Dong Xin[4]提出的个人信息集成与管理平台 SEMEX 系统采用以数据为主的体系架构,采用数据源、域模型、关联与实例、领域模型和关系抽取引擎来组织管理数据。J P Dittrich[6]提出了一个基于图数据模型和资源视图的 iDM 模型,用一种统一资源视图的概念和形式化表示方法,实现各种数据类型(如文档、目录、关系表、XML 文档、数据流等)的统一表示,采用数据源层、个人数据空间管理系统 PDSMS(Personal DataSpace Management System)、应用层的分层体系架构来组织管理数据。

中国人民大学的孟小峰教授发表了关于数据空间技术发展的综述性文章,并提出了一个典型的数据空间集成与管理框架,该框架由数据集成引擎、数据空间引擎、数据演化引擎和数据输出引擎组成[1]。钟鸣等人基于 RDF 提出了类似的元组模型,采用逐层分解的方式构建图,并提供了强大的查询能力[7]。董彦磊等人提出了一个应用于数据空间的 3 层组织结构,该结构由物理数据层、逻辑数据层和应用层组成[8]。逻辑数据层是整个数据空间的关键组成部分,基于该层才能对数据空间进行统一的管理,同时支持数据空间管理系统所提供的各种服务。杨丹等人以实体作为基本的数据单位,提出分层的图模型 lgDM[9],用来建模数据空间中存在的各种异构数据,即:实体关联数据图和实体关联模式图。王江海等人基于刻面的概念,利用数据源、刻面和属性来描述数据源[10]。

李玉坤等人针对数据空间本质特征,提出了基于图的个人数据空间概念模型和基于四元组的数据空间逻辑模型,该模型可以刻画数据空间的时序特征[11]。概念上将个人数据空间用一个大的有向图表示,图中节点表示数据对象,边表示数据对象之间的关联关系,数据对象和关联可以具有若干属性,属性取值具有时间属性。逻辑上用四元组<对象,属性,取值,时间>描述个人数据对象及其动态变化,即用形如<Object, Attribute, Value, TimeStamp>的四元组来刻画个人数据空间,其语义表示为"一个数据对象的一个属性在特定时间的取值"。刘正涛在数据空间的基础上,进一步提出了一种新的 Web 数据管理方法,即 Web 数据空间[12]。通过 pay-as-you-go 的构建方式,利用语义集成 Web 上的数据访问,实现一个 Web 数据集成系统,此系统的特点是可持续改进性,系统为组织或个人提供了一种有效利用 Web 数据的途径。

企业数据空间的数据是复杂多样的,现有的个人数据空间的数据组织方法或者传统的数据组织方法都无法将"先模式后数据"和"先数据后模式"的两种数据管理方式灵活地结合,且企业中的半结构化数据和非结构化数据越来越重要,需要统一的方式对企业中的异构数据进行统一表达和描述,且随着企业业务中的快速变化,需要一个灵活的企业数据模式的描述方式,可以随着企业的变化,更好地满足企业中的数据应用需求,需要对企业数据空间的数据组织方法进行进一步的研究。

2　企业数据空间的分层组织模型

2.1　企业数据空间概念

个人数据空间管理的主要是与个人相关的数据,仅需满足个人数据需求即可,而企业数据管理的对象是整个企业中所有相关的数据,需要满足各种应用系统的数据服务需求。与个人数据管理相比,企业数据管理更加复杂,结合数据空间的概念,提出企业数据空间的概念:

定义 1:企业数据空间(Enterprise Data Space, EDS)是以整个企业为主体,以企业中各个部门的信息系统中的数据和数据间的关联关系为管理对象的数据空间,提供按需、即时、灵活的数据服务[13]。

企业数据空间的主要特点有:

(1)可以对结构化、半结构化和非结构化的数据进行统一描述和管理。

(2)将原有关系数据库的"先模式后数据"和数据

空间的“先数据后模式”的方式协调起来进行管理,灵活管理两种方式的数据。

(3)具有多维度、多层次、多角度的数据组织方式,更能满足企业灵活管理和使用数据的需求。

2.2 分层组织模型

为了更加灵活和高效地组织管理企业数据空间中海量的多源异构数据资源,该文提出了一个应用于企业数据空间的分层数据组织结构,按照数据资源目录、数据模型、数据三个层次进行组织与管理,如图 1 所示。

其中数据空间(DataSpace,DS)是与主体相关的所有数据和数据间关系的集合。企业中不同的部门、项目组或者个人,都可以根据需要创建数据空间,并对其进行维护和使用。不同数据空间中的数据也可以存在交叉,可以看作企业数据空间的个人视图。

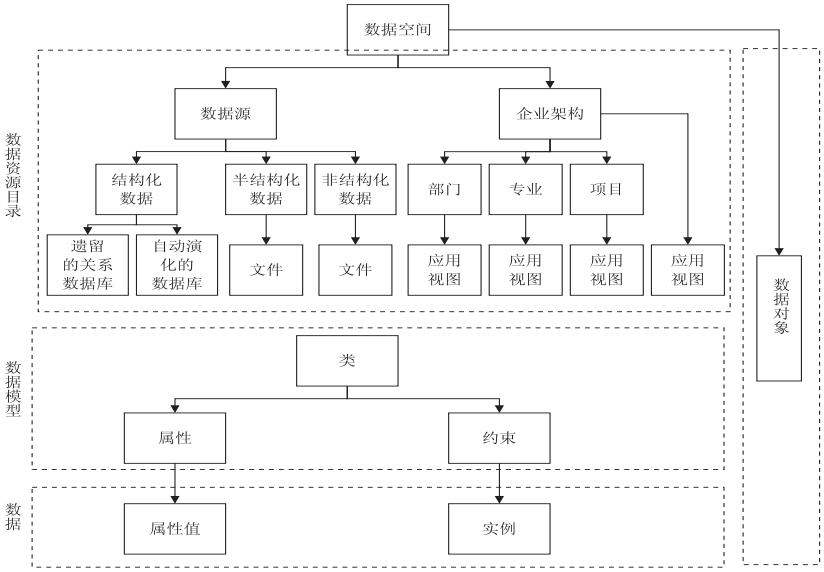

图 1 企业数据空间的分层组织架构

数据资源目录(Data Resource Catalog,DRC),是从多维多角度对数据空间中的数据进行分类和组织的一种树形目录结构,也是企业中数据的分类标准。数据源是从数据存储载体形式进行分类的,企业架构是从使用数据的角度对数据进行分类,采用应用视图的方式使用数据。这两种分类方式是数据空间提供的默认分类方式,企业根据需要可以自行定义相应的分类角度。数据资源目录的每一个叶子节点都对应有相应的数据资源,数据资源可以是实际的数据源,如具体的数据库系统、文档文件等,也可以是虚拟的数据源,如数据视图。在数据空间中,数据资源目录的结构是灵活的、动态的,一个数据资源可以属于多个目录节点。

数据模型(DataModel,DM),代表着不同数据资源的数据结构,包括多种类型的数据源模型,既有物理数据源的存储模型,又有虚拟数据源的逻辑模型,还包括数据空间的逻辑数据模型,即企业数据空间的所有数据资源作为企业顶层组织的一个数据视图。数据模型有两种情况:一种是遗留的关系数据库的数据模型,按照“先模式后数据”的形式,将关系数据库的数据模型直接纳入企业数据空间的管理之中;另一种是,预设数据模式之外或没有严格数据模式的数据,随着企业的需求改变和数据模式演化,逐渐演化出来的数据模式。数据模型由类、类之间的关系、数据操作和数据约束组成。类(Class),代表着数据模型中的类(也就是实体),例如关系数据库的数据模型中的一张表,半结构化数据中的元素。属性(Attribute),代表着类中的属性,如关系表中的字段,半结构化数据中的元素。约束(Constrain),代表数据模型中类的相关约束。

数据(Data),是指符合数据模型定义的类的数据实例,即数据源。企业数据空间的数据源分四类:遗留的关系型数据库、半结构化的数据文件、非结构数据文件和自动演化的数据库。其中遗留的关系数据库指的是,企业中已经投入使用的各个信息系统或者应用的关系型数据库,其中包含大量数据模式稳定的结构化数据;半结构化数据文件主要指的是 XML 文件;非结构化数据文件常见的有文档、视频、音频和邮件等;自动演化的数据库主要指的是,事先没有建立完整的数据模式的数据,而且也没有相应的物理存储模式,直接将数据存入 ESD 空间中,通过数据模式演化,自动创建相应的逻辑模式和相应的物理存储模式,提供该数据模式下数据的存储和管理。

数据对象(Data Object),是指没有相应明确数据模式和结构的数据,直接纳入企业数据空间的管理,随着企业数据空间的使用或者有需要的时候,逐步完善数据对象的数据模式,形成更加严格的数据模式,利用数据模型表示出来,进行组织和管理。

3　基于属性图模型的数据模型描述方法

3.1　属性图数据模型

企业数据空间中的数据结构复杂,包含不同异构数据的数据模型,所以需要一个可以描述企业数据空间中所有数据的方法。该文利用属性图数据模型(Property Graph Data Model,PGDM)[14]来描述数据空间中存在的各种异构数据。企业数据空间利用属性图模型将所有数据描述并关联起来,形成一个与企业相关的属性图。其中使用的基本概念定义如下:

定义 2:节点(Node)是属性图模型中的一个基本元素,用来表示各种类型的数据,可以是数据源、数据资源目录分类节点,数据模型中类、属性、约束,数据层的每一个数据单元,数据对象等。节点的标签(Label)表示数据的类型或模式信息,属性集(Properties)描述节点的具体信息,节点可以包含多个属性(Property)和多个标签(Label),每个节点至少拥有一个用于区分节点和节点之间是否相等的唯一标识。

定义 3:关系(Relationship)是任意两个节点间可能存在的关联关系,同样是属性图模型中的基本元素,将节点关联起来构成图,也可以称为图论中的边(Edge)。其始端(Start node)和末端(End node)都必须是节点,关系不能指向空也不能从空发起,而且关系是有方向的。关系和节点一样可以包含多个属性,但关系只能有一个类型(Type),一个节点可以被多个关系指向或作为关系的起始节点。

定义 4:属性(Property)是节点或者关系所具有的特性,节点和关系都可以有多个属性。属性是由键值对<key,value>组成的,就像 Java 的哈希表一样,属性名类似变量名,属性值类似变量值。属性值可以是基本的数据类型,或者由基本数据类型组成的数组。

定义 5:节点标签(Node Labels)是一种对节点进行语义分类的方法[14]。节点可以分配零个标签、一个或多个标签,标签本质上是图形结构中面向集合的概念:它们允许轻松高效地创建子图,这对于许多不同的用途非常有用,例如仅查询数据库内容的一部分。可以使用标签表示某种数据类型、结构或模式,或者根据企业需要,自定义相应的标签。虽然不是必需的,但节点应至少具有一个标签,为了更加清晰地了解数据。

定义 6:关系类型(Relationship Types)实现的内容与处理节点标签类似,是为了对关系进行分类。但是关系类型是关系必不可少的,每个关系必须有一种且只有一种类型,两个节点可以由多个关系连接,并且在属性图中复杂、深层遍历期间使用。

定义 7:属性图模型的数据结构,可以形式化定义为一个二元组 PGDM = (Nodes, Relationships),其中 Nodes 表示企业数据空间中所有的节点集合;Relationships⊆Node×Node 表示节点之间的关系集合,具体情况如图 2 所示。

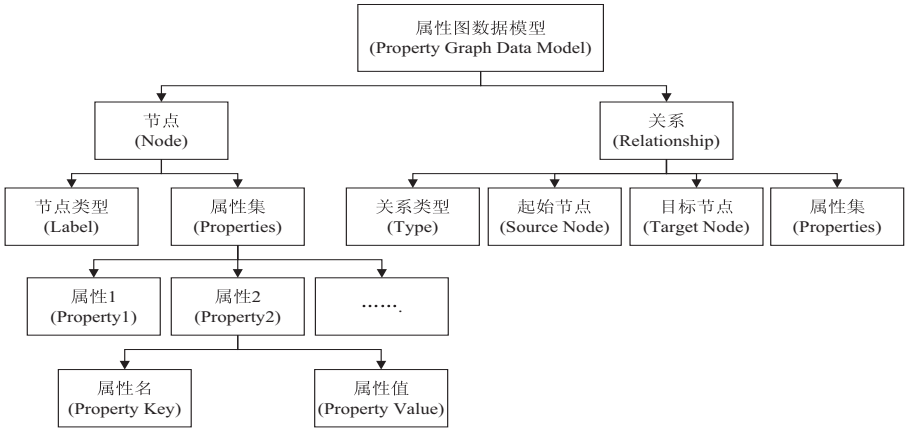

图 2　属性图数据模型的数据结构

利用属性图数据模型对企业数据空间中的异构数据进行统一描述,具有以下优点:

(1)属性图模型没有固定的模式。属性图本身并不强制要求数据必须具有严格的关系模式,符合数据空间先模式后数据的特点。

(2)属性图的无模式和弱模式特性,更适合处理半结构化数据。当半结构化数据发生变化的时候,可以比较方便地处理数据模式的变化。

(3)节点、属性及关系,更符合现实世界中事物的特点,更加容易理解。

(4)关系是明确的,不是由某种约束推断的,也不是通过连接操作在查询时建立的,是属性图中重要的基本元素,而且可以具有属性,可以将现实世界中数据间复杂的关系给描述和利用起来,具有和节点相同的表达能力。

(5)独立于各种数据源,可以描述结构化、半结构、非结构化数据结构,可以多维度多层次描述企业数据空间数据,便于自动模式演化。

3.2　异构数据模型的描述方法

为了将企业中大量稳定的关系数据库也集成到

EDS 中进行管理,将“先模式后数据”的特点也体现出来,还有后续弱模式或无模式的数据对象随着使用,逐渐演化出相应的数据模型,设计统一的数据模式描述方法,无论是结构化数据模型、半结构化数据模型或者 ESD 全局逻辑模型,都用同一种描述方法进行描述,在逻辑模式层上进行统一管理。

数据模式由类、类之间关系和类的约束组成。用来统一描述企业数据空间中的各种模型和企业数据空间的概念模型。描述方法的形式化表达为:

DataModel = {Class, Constrain, Relationship}

其中,DataModel 代表不同的数据模型,既有物理数据源的存储模型,又有虚拟数据源的逻辑模型。在企业数据空间中,有一个全局的逻辑模型。企业数据空间的所有数据资源作为企业顶层组织的一个数据视图,其对应的数据模型即全局逻辑模型。

Class 是数据模型下包含的各个类(也可叫做实体),例如关系数据库中的一张表或者是视图,或者一个半结构化的文件的元素节点。其中 Class 下包含不同的属性(Attribute),如关系表中的字段,半结构化数据中的属性节点。

Constrain 是类的相关约束,是对类的约束限制,如关系表的字段的取值约束。

Relationship 是数据模型下各个类之间的关系,如常见的有关系表中的主外键关系、类与类之间的引用关系等。

通过上述的数据模型描述方法,无论是各种数据源的数据模型,还是虚拟的数据模型或者数据空间本身的全局模型,都可以用统一的方法描述。不同异构数据模型,用同一种数据模型描述方法,为后续数据模式匹配提供了良好的基础,可以更好地支持后续的数据模式演化。

3.2.1 描述结构化数据模型

当结构化数据模型在 EDS 中被进行描述时,需要有一定的描述规则,才可以直接地对关系数据结构进行描述。其中的描述规则有:

(1)其中的表(Relation)用 Class 下的节点来描述,表的一些本身特征用 Class 下节点的属性集来表示。

(2)其中表的各个字段(Attribute)用 Attribute 来表示,属性的本身特征用 Attribute 下的属性集来表示。

(3)表的一些完整性约束条件和用户自定义的约束条件,用 Constrain 来描述。

(4)表与表之间的关系,用 Relationship 来描述。

3.2.2 描述半结构化数据模型

半结构化数据的数据结构和数据内容是混合在一起的,介于结构化和非结构化数据之间。现在企业中用的最广泛的半结构数据,就是 XML 文件。以 XML 的数据模式为例,其数据模式描述方法为:

(1)将 XML 文件中不含有文本节点的元素节点用 Class 下的节点来表达,元素节点的名字当作类的名字,属性节点当作节点的属性。

(2)将 XML 文件中含有文本节点的元素节点用 Attribute 下的节点来描述,元素节点的名字当作属性的名字,属性节点当作节点的属性。

3.2.3 描述非结构化数据模型

非结构化数据常见的有视频、音频、文档或者一些二进制文件,没有明显的数据结构。但是在非结构化数据文件中,其实是隐含着相应的数据结构的。如常见的音频数据、视频数据,WORD、PDF 文档,石油企业中的地震数据、测井数据等,这些数据的格式是标准化的,格式的描述不在数据体中。贲福才提出了一种非结构化数据模式描述标记语言 BULKML,该标记语言采用 XML 描述的非结构化数据的结构,为非结构化数据补充模式描述,使非结构化数据转换为半结构化数据[15]。BULKML 按数据文件偏移量(二进制文件)或文件标记(文本文件),对数据文件中的数据的语义进行标注。BULKML 按数据文件格式规范进行定义,每一种格式规范定义一个 BULKML。而且在国家标准非结构化数据表示规范中,利用 XML 文件格式来表示非结构化数据文件的数据结构[16]。也就是说非结构化数据的数据模式用半结构化数据文件来表示,从而利用半结构化数据的数据模式描述方法,实现对非结构化数据的数据模型进行描述。

4 结束语

该文以企业数据的现有数据管理特征为出发点,针对灵活高效地组织企业数据空间中的数据资源的问题,对企业数据空间的数据组织方法进行研究:提出了企业数据空间分层的组织模型,实现对企业数据空间多角度多维度的组织;建立了基于属性图模型的数据描述方法,统一描述各种异构数据。利用此方法可以高效灵活地组织和管理企业数据空间的数据,为后续的数据模式演化奠定基础。

在以后的工作中,将致力于改进企业数据空间的数据的存取优化,考虑将企业中的实时数据也纳入企业数据空间的管理,解决数据模式演化问题等,使企业数据空间功能更加完善。

参考文献:

[1] 李玉坤,孟小峰,张相於. 数据空间技术研究[J]. 软件学报,2008,19(8):2018-2031.

(下转第 65 页)