文章编号: 1002-0411(2001)06-498-04

数据仓库元数据的界定与分类

吕 波[1] 王延章[1] 王红梅[2]

(1. 大连理工大学信息与决策技术研究所 116024; 2. 吉林工学院计算机系 130022))

摘 要: 元数据是数据仓库研究的热点, 目前对元数据有多种定义方式, 如: 描述数据的数据, 数据和信息之间的桥梁等, 这些面向应用角度的定义, 并不能真正表达元数据的内涵. 本文从元数据和基础数据分离过程入手, 研究元数据的定义和分类, 对元数据的概念内涵进行深入探讨, 提出广义元数据和狭义元数据的概念, 扩展元数据的定义, 克服面向应用角度定义元数据的不足.

关键词: 元数据; 数据仓库; 广义元数据; 狭义元数据

中图分类号: TP311 文献标识码: B

DEFINFITION AND CLASSIFICATION METADATA FOR DATA WAREHOUSE

LU Bo[1] WANG Yan-zhang[1] WANG Hong-mei[2]

(1. *Institute of Information and Decision Technology, DUT*; 2. *Jilin Institute Of Technology*)

Abstract: As a hotspot of data-warehouse, there are several methods to define metadata, such as data about data, a bridge between data and information, these definitions were put forward, depending on different applications, and they can not express the significance of metadata. The paper began from studying the process that metadata was departed from underlying data, then the classification and definition of metadata were studied, at last we came up with the metadata in general sense and the metadata in narrow sense, these definitions overcome the shortcomings of the definition oriented special applications.

Keywords: metadata, data-warehouse, metadata in general sense, metadata in narrow sense

1 引言(Introduction)

由于企业中广泛存在“数据监狱”和“数据贫穷”现象, 致使决策者无法得到企业的全局决策信息和宝贵的历史信息, 为解决这一矛盾现象, 一种新兴的数据存储和处理技术—— 数据仓库技术, 应运而生. 数据仓库技术需要解决诸如: 数据跨平台, 大量数据(数百 GB 数据)的有效存储问题, 于是元数据的概念被引入到数据仓库中. 元数据的研究吸引了众多研究者的关注, 但多数的研究工作局限于从特定的应用引出元数据的分类, 文[2]从元数据获取和元数据存取的角度提出前仓元数据(front room metadata)和后仓元数据(back room metadata); 文[3]从数据仓库的组织和使用的角度把元数据分为管理元数据和用户元数据, 文[4]中又提到技术元数据、商业元数据和信息浏览元数据, 此外, 还有部分研究工作具体地探讨某一类型元数据, 如文[5]研究质量元数据(quality metadata)和寿命元数据(longevity metadata); 文[6]研究地理空间元数据. 目前, 缺少对元数据比较全面的分类研究和元数据概念内涵的深入剖析.

从以上的分析看出, 目前对元数据研究不够深入, 其根本原因在于缺少对元数据概念内涵的深入研究, 虽然有些文献[2, 3]尽可能的列举了所有的元数据, 但没有给出元数据的严密定义, 本文从元数据的发展状况和现有元数据的类型研究入手, 提出广义和狭义元数据的定义, 这种全新的元数据定义方法, 拓展了元数据的概念内涵. 本文下面几部分的组织如下: 第二部分研究元数据和基础数据分离的意义, 第三部分研究数据仓库中元数据的分类, 第四部分提出广义元数据和狭义元数据的概念并对这种定义方法进行分析, 最后给出基于这种定义的后续研究内容.

收稿日期: 2000-12-12
基金项目: 国家自然科学基金资助项目: 编号 60074038

2　元数据和基础数据分离(Separation of Metadata and Basic Data)

元数据和基础数据分离的过程贯穿从文件系统到数据仓库技术的发展历程. 早期的事物 i 处理系统数据和程序结合在一起, 随着数据量的增大, 程序和数据相互依赖的不灵活性表现得越来越严重, 致使程序的开发和维护的费用高, 于是出现了数据和程序的分离. 数据单独存放在数据文件中, 在一定程度上实现了元数据和基础数据的分离, 这时的元数据和程序结合在一起, 用来定义数据结构、完成数据存取操作和有效管理数据文件的功能的程序代码.

通过元数据管理数据文件, 虽然在一定程度上解决了操作灵活性的问题, 但存在数据冗余、数据安全性差等问题, 此外, 尽管程序和基础数据实现分离, 但还和元数据结合在一起, 数据结构或存取文件的任何变化都将引起程序代码的改变. 60 年代末出现的数据库管理系统(DBMS), 解决了传统文件组织所产生的问题: 数据冗余、并发操作, 实现了元数据和程序的分离.

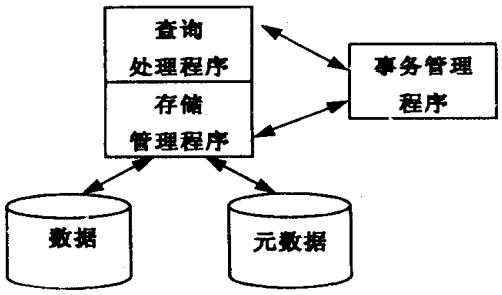

图 1　元数据是 DBMS 的主要组成部分

Fig. 1　Medata is the important part of DBMS

为对单独存放的数据文件中的数据直接进行查询、管理, 实现数据重用、共享, 发展了数据库的概念, 各种数据库管理系统(DBMS) 能够有效的存取和操作大量数据. DBMS中元数据描述数据结构的信息, 如实体/关系模型、属性名、属性的数据类型等信息. 在 DBMS 中, 元数据单独存储在数据文件中, 如 Foxpro 中扩展名为 ". CDX" 的文件存放元数据信息, 在这一阶段, 元数据的结构相对简单, 功能比较单一.

随着数据库技术发展, 尤其是关系数据库模型理论的成熟, 各商家纷纷推出自己的数据库管理系统. 这些数据库管理系统在企业中得到广泛应用, 由于各企业甚至是同一企业的不同部门所采用的数据库平台存在很大差异, 所定义的数据格式和编码方式各异. 在这种环境下, 企业决策者要想得到企业全局的决策是十分困难的. 另一方面, 为了进行有效的决策, 决策信息不仅应该是全面的还应该是完整的, 即决策者不仅需要当前的数据, 还需要过去的历史数据, 才能完成各种复杂分析, 如趋势预测和数据挖掘, 以支持决策, 这种需求导致 OLTP 系统和 OLAP 系统及其支持环境的分离, 一种新型的数据存储和处理技术—— 数据仓库产生了.

William Inmon 提出数据仓库的描述: "一个数据仓库通常是一个面向主题的、集成的、随时间变化的、但信息本身相对稳定的数据集合, 它用于对管理决策提供支持." Inmon 的定义概括了数据仓库的几个特点: 面向主题的、集成的、稳定的、历史数据的集合.

随着数据仓库技术的不断发展, 元数据在数据仓库中的作用日益显著. 元数据不仅定义了数据仓库的作用、指明了数据仓库中信息的内容和位置、刻画了数据的抽取和转化规则、存取了数据仓库的主题和相关信息, 而且实现了数据仓库的管理, 如修改和跟踪数据、描述数据同步需求衡量数据质量等功能. 按一定格式组织的元数据可以方便、快捷地实现数据的存取、转换、分析、管理和分布数据的共享, 成为整个数据仓库的管理核心.

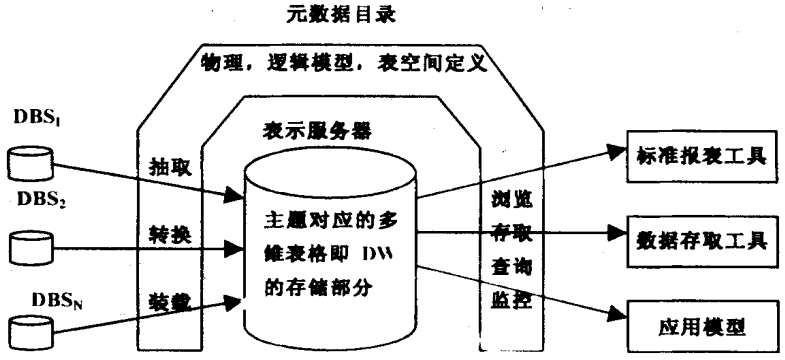

图 2　元数据在数据仓库中的作用

Fig. 2　Metadata's function in data warehouse

图 1 给出了数据仓库中元数据的基本功能, 和 DBMS 中元数据所起的作用相比, 它不仅功能更加强大, 所存在的位置也具有离散的特点, 更好的管理和维护元数据成为数据仓库技术的重要课题, 下面通过分析当前元数据的分类和定义提出新的元数据定义方法.

3 元数据分类(Classification of Metadata)

在数据仓库中元数据无处不在, 除了基础数据本身就是元数据, 甚至把元数据比作数据仓库的 DNA, 它定义了其它元素的功能及属性. 随着数据仓库理论研究的不断深入, 研究者越来越重视元数据概念和分类的研究, 目前提出的元数据的分类方法中较具代表性的是文献[2]中提出的: 前仓元数据, 后仓元数据. 这种分类方法是从数据仓库中数据的存取和数据获取的角度体现元数据的功能.

3.1 后仓元数据

在数据仓库中实现数据获取的元数据就是后仓元数据, 数据获取就是适时地把数据从源系统通过一定的转换送到表示服务器的过程. 可见, 后仓元数据与过程相关, 负责管理和引导数据的抽取、清洗及数据的装载过程. 后仓元数据又可进一步分为: 源系统元数据、数据进阶元数据及 DBMS 元数据. 数据仓库中数据的来源非常广泛, 可以是各数据库平台、电子表格、各种无格式或有格式文档甚至是一些在线资源, 源系统元数据实现对这些数据源的存储介质、数据模式的说明; 同时, 各数据源属于不同的部门, 具有特定的商业内涵, 被定期刷新和维护, 用户只能按一定的权限存取数据源信息, 这些都是源系统元数据需要描述的内容. 源系统元数据是对源系统的标识, 为数据进阶准备信息.

数据进阶区是数据仓库的工作平台, 在这里原始数据被抽取、清洗、组合、装载、存档并快速的输出到一个或多个表示服务器平台上. 进阶元数据控制了整个进阶过程, 依据数据进阶的不同阶段: 抽取、转换、装载, 进阶元数据要定义的内容有以下几点: 描述数据获取工作流, 根据原始数据组织方法及数据仓库的需求选择抽取工具, 并对抽取时间、抽取内容、及数据完整性给预说明; 定义数据转换表和域映射, 给出记录筛选规则; 说明数据装载的方法(批量装载或单记录载入方式); 为保证数据质量, 进阶元数据还要说明数据质量审计指标及参数, 记录进阶日志等. 数据进阶工作相当复杂, 工作量占数据仓库开发的 80% 以上.

一旦源数据传送到数据仓库或数据集市的 DBMS 中, DBMS 元数据开始起作用. DBMS 元数据描述数据库的逻辑结构和物理组织, 提高查询效率、实现查询优化, 保证数据存取的安全性和一致性. 建立在关系模型基础之上的数据仓库的 DBMS 分区信息、索引信息、视定义、存储过程 SQL 管理脚本、DBMS 备份状态、备份过程、备份安全性.

当数据最后装载到表示服务器上, 要为用户提供友好、安全地存取功能, 就需要前仓元数据.

3.2 前仓元数据

在数据仓库中实现数据存取的元数据是前仓元数据, 数据仓库的前端是数据仓库公众形象, 它提供一系列的服务, 从而使用户以尽可能简化的方式获得数据仓库信息. 按数据仓库前端所提供的服务, 前仓元数据描述以下内容: 内容简化元数据, 包括表、行、组群的描述, 即商业结构, 预定义的连接、联合规则, 复杂用户查询公式的重构和重定向, 例如: 上一年的销售额增长数量, 对于基于 WEB 的前端还要描述网络安全、用户验证、访问日志, 数据元素, 表、视图、的存取映射, 资源费用统计.

采用这种元数据分类方式, 虽然能够给出元数据完整的列表, 但无益于元数据的表示与管理, 为此本文提出了一种全新的元数据定义方法, 提出广义元数据, 狭义元数据的概念.

4 广义元数据和狭义元数据(Metadata in General Sense and Metadata in Narrow Sense)

目前对元数据的定义大多是面向特定的应用, 尤其是数据密集型应用, 如: 地理信息系统、多媒体系统、数据挖掘系统, 面向应用的元数据定义并不能刻画元数据的真正内涵. 本文对已有的元数据的概念进行扩展, 提出广义元数据和狭义元数据的概念.

元数据首先是数据, 在汉英辞典上对数据的定义: 科学实验、检验、统计等所获得的和用于科学研究、技术设计、查证、决策等的数值, 也就是, 从狭义上讲数据是一些数值, 而在计算机科学中数据在最基本的层面上也是数值, 按各种方式组织的数值在不同层次上的理解可以是字母、文字、其他特殊字符, 也可以是图形、图像、声音等多媒体数据, 而在更高的层次上对这些数据操作的程序、文件即除去计算机硬件之外的都是数据或数据的集合, 这就是广义的数据. 无论是广义数据或狭义数据只有通过一定的组织管理才能成为真正有用的信息, 这就需要

元数据这个桥梁, 元数据的桥梁作用可以表示如下:

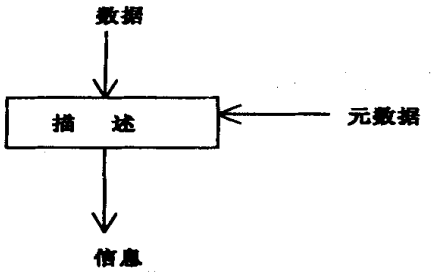

图 3(a)　描述型元数据的作用

Fig3.　(a) Descriptive Metadata

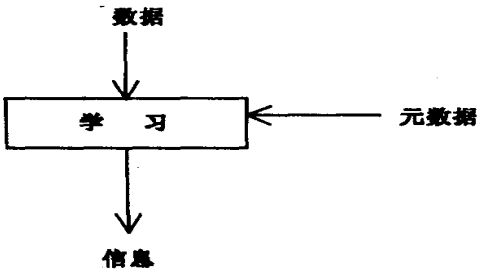

图 3(b)　知识型元数据的作用

Fig. 3　(b) Knowledge Metadata

所以元数据不仅具有对基础数据的描述作用, 例如: 描述数据或数据集的类型、值域、所用者、数据的刷新频率等(如图 3a 所示); 而且, 知识型、操作型元数据还可以使基础数据通过学习成为支持决策的信息, 例如: 知识型元数据时序指标分析模型、线性规划算子可以对一组数据分析, 所得分析结果可以用各种图形, 如饼图、直方图等表示, 成为决策依据(如图 3b 所示).

从前面对元数据与数据分离过程及数据仓库中元数据分类的分析可以看出, 元数据是一个特定层面的数据, 它使存储数据成为可以被利用的信息, 或者说元数据是信息和数据之间的一座桥梁, 下面给出广义元数据和狭义元数据的定义:

定义 **1**　狭义元数据

描述数据的内容、质量、环境及数据的一些其他特色的数据叫做狭义元数据.

定义 **2**　广义元数据

实现对数据的描述、转换、操作、管理的数据和知识称为广义元数据, 广义元数据在不同层次上对数据进行表示、管理和操作, 它可以是关系映射、具有输入输出接口的算子和模型、用谓词表示的知识等.

我们可以把这种元数据定义方法用下面的关系图表示, 如图 4。

元数据应建模为知识[10], 或从知识结构的角度描述元数据, 广义的知识即广义的元数据应该包括基础数据和数据之间的关系, 基础知识和描述知识的知识. 其中基础数据是所有信息的最终表示形式和实现依据.

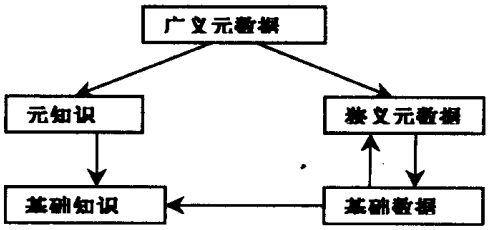

图 4　元数据定义关系图

Fig. 4　Relation chart for metadata definition

这里提出知识和元知识扩展狭义元数据的概念, 知识是在一定范畴的知识, 按照数据仓库实现数据融合、数据分析以支持决策的要求, 这里所定义的知识和元知识又可以按照下面的层次分为四类:

1) 客观事物系统知识, 它是问题所涉及的客观事物, 或称对象系统的知识, 一般也称为问题的领域知识, 主体是事物的的概念、属性、相互联系机制和结构知识.

2) 问题知识, 它是问题定义和描述的知识. 这种知识主体是人的主观目标要求方面的知识和涉及的客观事物的概念和关系的知识.

3) 方法性质是, 它是关于问题的一种同态映射的知识, 主体上是问题的广义模型, 如投入产出、线性方程、状态方程和经济计量学模型等分析性数学模型, 或线性规划、非线性规划、动态优化、最优控制等优化模型. 更一般的包括模糊的、随机的、尤其是经验性模型等.

4) 工具性或算子类知识, 这是一种高度形式化的知识, 主体上是通用的算法或解题器类知识, 它包括从简单的矩阵算子到单纯形迭代法、梯度迭代法、动态规划法、多目标优化方法, 以及逻辑推理和启发式搜索算法等程序或软件包.

其中知识 3), 4) 属于基础知识是形式化的可以独立于问题领域, 而 1), 2) 对不同的应用具有领域的特殊性, 对形式化的知识进行描述, 以支持具体的决策问题.

这种元数据定义方法是元数据理论和数据仓库开发实践相结合的结果. 在刚刚接受验收的九五攻关项目“国民经济辅助决策 GIS”的研究, 完全建立在综合数据仓库开发平台之上. 在系统中不仅实现基础数据和狭义元数据分离, 而且基础知识和元知识的分离, 其中基础知识是积累多年的各种模型和

(下转第 507 页)