基于元数据的异构专家信息管理研究①

Research of Heterogeneous Expert Information Management Based on Metadata

刘高嵩 麻广伟 龙 军 (中南大学 信息科学与工程学院 湖南 长沙 410075)

摘 要: 元数据在网络信息资源的管理、存储和检索中发挥着重大的作用。本文对元数据和基于元数据进行管理的理论基础进行了探讨,说明了元数据的内容、结构、生成过程和依据元数据的管理模式,采用 XML 描述元数据标准,提出了一种以元数据为核心,以元数据驱动任务操作的异构数据管理策略。将其思想加以扩展并抽象为一个框架模型应用到专家信息管理中,在元数据的层次上来定义和管理异构专家信息。

关键词: 元数据 管理 异构数据 专家信息

1 引言

随着计算机网络的发展和知识经济时代的来临,人力资源已成为生产力发展的重要因素,各行业专家人才信息的管理一直是经济社会发展的热点课题。近些年来,全国各部门、各省市都建立了各自的专家库,积累了大量的专家信息,然而在长期的数据积累过程中,各部门各地区各自为政,信息资源标准并不统一,从而导致不同部门之间的数据协作困难重重,不同种类的专家数据也难以管理。

因此,异构专家信息的有效管理已刻不容缓,一方面,专家数据生产者需要有效的数据存储和维护方法,专家信息必须有统一的数据格式来描述;另一方面,用户需要有效的访问数据集以获取所需数据、进而产生出所需专家信息的方法。传统的管理方式已无法有效的存取和利用这些数据,为了满足各方面对专家信息的管理和应用要求,迫切需要对原始专家数据进行描述的信息—元数据(metadata)。

2 专家信息的元数据分析

2.1 元数据的定义

元数据是关于数据的信息,是具有描述、解释、定位信息资源功能的结构化信息[1],是说明数据内容、质量、状况及其他有关特征的描述信息[2]。它在专家信息系统中用于描述专家数据集的内容、质量、表示方式、生成时间、管理方式以及数据集的其它特征,有了元数据可以使得获取、使用和管理信息更加容易,它极大的方便了异构数据管理。

2.2 元数据的内容和格式

元数据的内容从本质上讲是对数据文档特征(资源)所做的描述,元数据基本内容的抽象形式为<数据文档,特征,特征值>,通常称之为元数据元素,较高的层次的内容还有元数据实体、元数据节,这 3 个层次的定义[2,3]如下:

(1) 元数据元素:元数据元素是元数据最基本的单元,它描述原始数据某一个特征,按照数据库语言,它是填入数据的“字段”,任何元数据最后都是直接或者间接(通过元数据节、元数据实体)由元数据元素来描述。

(2) 元数据实体:若干个相关的元数据元素构成元数据实体,它描述原始数据某一方面的若干个特征。

(3) 元数据节:元数据节是由相关的元数据节、元数据实体和元数据元素构成的元数据的子集。

2.3 描述专家信息的元数据标准

元数据标准规定了元数据节、元数据实体、元数据元素这些元数据内容在元数据中的组织规则,它可以采用 XML Schema 等来表现。XML Schema 是一

① 基金项目:国家自然科学基金(60873081)
收稿时间:2008-10-27

种用来描述 XML 文档， 控制文档结构的方法。作为 XML 语言的主要模块，Schema 对标识的标准化起着极其重要的作用，元数据是按照元数据标准(XML Schema)的规定生成的信息文档(XML 文件)[4,5]。

专家信息涉及到学科、学历、学位、单位、党派、民族、职称、国籍、社会兼职、学术成果等相关信息。综合来看，专家信息元数据标准[6]内容分两个层次：第一层是目录信息，主要用于对数据集信息进行宏观描述，它适合在管理和查询专家信息时使用；第二层是详细信息，用来详细或全面描述专家信息的元数据标准内容，是数据集生产者在提供空间数据集时必须提供的信息。

为此，可以把描述专家信息数据集的元数据分为三类：即模式信息、定位(导航)信息和其他相关信息。模式信息是用于描述专家数据集的数据结构和语义信息；定位元数据提供了如何得到一个数据源的信息；其他相关的元数据为数据源提供附近的描述信息。

3　专家信息的元数据生成过程

同一个数据在不同的应用领域会有不同的元数据来描述它，本文讨论的管理要求原始数据具有同领域的性质，这样就可以保证按照该领域的统一的元数据标准来生成元数据，从而为管理提供一个统一的逻辑视图。本文用数据文档来表示数据集中的数据单元，那么，原始数据集由大量的数据文档构成；用信息文档来表示元数据集中的信息单元，那么，元数据集由大量的信息文档构成。本质上，在原始数据具有同领域的性质的条件下，原始数据集中某个数据单元所对应的元数据是一个描述该数据单元的有关特征的信息单元，数据单元和描述它的信息单元之间是一一对应的。

大量的异构数据文档只是机器可读的，是数据层次上的概念，它们以某种方式存在于各个原始数据集中，难以统一进行管理。在领域知识和常识的指导下，依据特定领域元数据内容的选取原则，领域专家们给出关于专家信息数据文档的有关特征的统一描述框架——描述专家信息的元数据标准(知识)。

按照特定领域的元数据标准[7]制定的描述专家信息的数据文档的有关特征的集合为：{特征 1，特征 2，…，特征 p}。数据文档的特征集合中的特征描述原本也只是机器可读的数据，但经过元数据标准和相关知识的处理后，获得{<数据文档，特征 1，特征值 1>，<数据文档，特征 2，特征值 2>，…，<数据文档，特征 p，特征值 p>}，把{<数据文档，特征 1，特征值 1>，<数据文档，特征 2，特征值 2>，…，<数据文档，特征 p，特征值 p>}按照描述专家信息的元数据标准规定的方式展现出来，这就成为了机器可理解的信息——元数据。元数据的生成过程如图 1 所示。

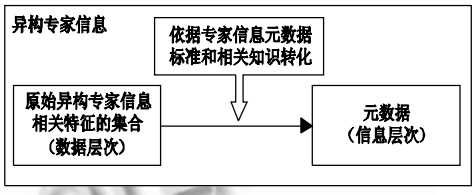

图 1　描述专家信息的元数据生成过程

这个生成过程有两层含义：

(1) 在某个原始数据集中，如果没有对其中的数据文档进行过局部的、不统一的描述，那么必须按照描述专家信息元数据标准逐个地描述数据文档的每个特征。

(2) 在某个原始数据集中，如果对其中的数据文档进行过局部的、不统一的描述，那么首先要把这些局部的、不统一的描述转变成为符合专家信息元数据标准的描述(统一描述词汇，保留标准中要求的特征描述，去除标准中不要求的特征描述)，此时得到的只是元数据标准要求的全部特征描述的一部分，再按照描述专家信息相关元数据标准把缺失的部分特征描述补足。

由上述生成过程可知，元数据是机器可理解的资源描述，所以元数据是信息层次上的概念，并且，数据文档和描述专家信息的有关特征的元数据间有着一一对应的关系。

4　基于元数据的专家信息管理方案

4.1 基于元数据的专家信息管理模式

由于难以直接对异构的大量数据文档(原始数据单元)进行管理，那么首先根据描述专家信息统一的元数据标准生成描述这些数据文档的信息文档(可以视为数据文档的逻辑视图，而且由于使用了统一的标准，该逻辑视图对于管理而言是统一的)，然后通过信息文档来对原始数据进行管理，这种管理模式有效屏蔽了原始数据的异构性，同时有较好的可扩展性(即新的专家信息相关原始数据或原始数据集可以方便地加入到原有的基于元数据的管理系统中去)，如图 2 所示。

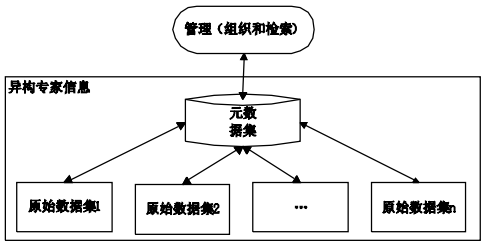

图 2　基于元数据的管理模式图

对管理中的组织和检索任务分别说明如下：

(1) 管理中的组织指的是根据信息文档把数据文档和信息文档按照一定的结构、顺序、排列方式组织起来。原始数据集和元数据集具体的组织形式可以多种多样，比如关系数据库、XML、文件系统等，最重要的是根据元数据标准生成的元数据与原始数据间要保持一一对应的关系[3]。

原始数据集＝原始数据集 1∪原始数据集 2!∪…，原始数据集 s={d_1，d_2 ，…，d_n}，其中 d_1 (1≤i≤n)是数据文档；元数据集={m_1，m_2 ，…，m_n}，其中 m_i(1≤i≤n)是元数据。元数据集中的元数据和原始数据集中的数据文档之间存在着一一对应的关系。

(2)管理中的检索[7]就是根据信息文档重新获得或恢复，是进行搜索、定位及读出数据文档和信息文档的过程。所有的检索任务均在信息文档集(统一的逻辑视图)上完成，只要检索出了满足用户需求的信息文档，那么根据这些信息文档、 数据文档与信息文档之间一一对应的关系，就能找出所有满足用户需求的数据文档，并将这些满足用户需求的数据文档和信息文档一起返回给用户。

4.2 异构专家信息的管理过程和说明

根据以上基于元数据的异构数据管理理论，引入统一的专家信息元数据标准，把各个专家信息源所做的局部的、不统一的、不规范的描述转化成为依据统一的关于专家信息的元数据标准生成的全局的、统一的、规范的描述(元数据)，在这个统一的逻辑视图上进行管理。结合图 3，模型管理过程如下：

(1) 各个专家信息数据源上报对各自所产生的数据的描述(即各个数据源各自的局部性的描述，这些局部性的描述是不统一、不规范的)，这些描述可能是文本、数据库纪录、XML 文件等。

(2) 首先利用同义词表按照专家信息相关标准统一那些局部性描述的描述词汇，然后元数据生成模块根据这些局部性的描述生成关于专家信息的元数据标准的 XML 文件，这些 XML 文件就构成了统一的、规范的元数据，同时，保留全局 id_expert(元数据的标识)和局部 id_local_expert(数据文档的标识)间的一一对应关系。

(3) 将元数据存入元数据库中。

(4) 从这些元数据 XML 文件中取出检索需要用到的部分专家信息目录，专家信息目录采用的是关系数据库表来组织。使用专家信息目录进行应用比直接在元数据库的 XML 文件上进行应用的效率要高；但专家信息目录实质上是元数据的一个子集，专家信息目录有时无法满足某些管理任务的要求，此时，就应该利用元数据库中存放的元数据。

(5) 管理模块的管理工作在异构数据文档的统一的逻辑视图(专家信息目录和元数据库)上来完成。

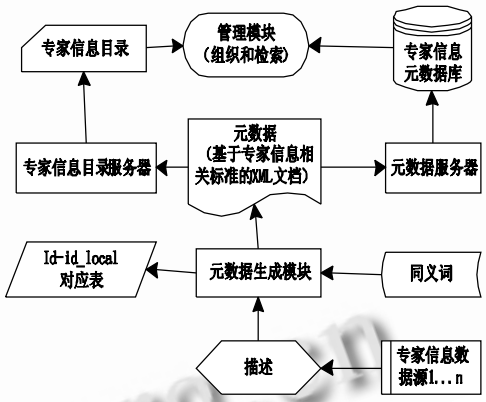

图 3　基于元数据的专家信息管理方案

4.3 元数据的管理模式

基于元数据的专家信息管理方案中必须涉及到对元数据的管理，元数据的管理主要由图 3 中的元数据服务器负责。元数据服务器的主要任务是当元数据服务的需要变化时，负责对任务的调度和分配； 还有当对元数据库中元数据进行添加和删除等操作时，对元数据库进行一致性检查等管理。

元数据服务器将具体的元数据操作封装成服务，当某子程序需要从元数据库中读取元数据时，元数据服务器首先调用元数据扫描服务对元数据库进行扫描， 若元数据库中存在相应的元数据，则调用数据读取服务对元数据库进行读取，然后经过元数据解析服务对读取的元数据进行解析，最后将解析后的数据传递给该子程序。当某子程序需要向元数据库中写入某

个元数据时，首先元数据服务器调用元数据扫描服务对元数据库进行扫描，若元数据库中不存在相应的元数据，则调用数据编辑服务对该元数据进行编辑，然后再调用元数据写入服务将该编辑好的元数据写入元数据库；若元数据库中存在相应的元数据，则元数据库管理器终止向元数据库中写入数据，并报告出错信息。当需求变化时，元数据服务器可以调用相应的服务对元数据库进行读取写入操作，体现出元数据模块的灵活性。元数据服务器架构如图 4 所示：

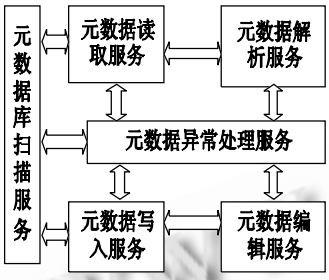

图 4　元数据服务器构架模型

5　总结

针对目前普遍存在的大量难以管理的异构专家信息的现状，本文引入元数据的方法来指导对这些数据的管理，阐述基于元数据的管理模式，提出了一种以元数据为核心，采用 XML 描述元数据标准，基于元数据的异构专家信息管理策略。本方案具有以下几个特点：(1)元数据的集中存储、集中管理模式，保证了元数据和数据集内容的一致性，能够有效地完成异构数据的管理和共享；(2)使用 XML 的异构数据融合技术，以专家信息核心元数据为基础，扩展了专家信息的数据描述功能；(3)采用了元数据的封装技术，将元数据的相关操作封装成服务，增强了模块的可扩展性。

参考文献

1 NISO Press. Understanding Metadata. 2004 http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.
2 应英.电子政务元数据管理系统的研究与设计[硕士学位论文].西安:西北工业大学, 2007.
3 徐财江,陈和平,陈志荣.土地利用现状数据元数据管理系统的设计与实现.2006 年中国土地学会学术年会论文集, 2006:707－713.
4 刘洪星.XML 建模和 XML 数据库建模.计算机科学, 2004.
5 陈海建,茅忠明. XML 本源数据库开放模型的设计与实现.计算机信息, 2006,4(3):238－240.
6 周骏,徐林,李征.元模型驱动的企业建模.计算机工程与应用, 2005,27(2):215－217.
7 张承伟,赖洪波,乌丽娟.政府信息资源元数据及其标准化的研究.计算机应用研究, 2006,12(1):51－53.