

# 基于数据元的数据集成技术研究

时贵英 文必龙 王志宝

(东北石油大学计算机与信息技术学院,大庆 163318)

**摘要** 针对当前数据集成方法只能精确到属性级映射的缺陷,提出了基于数据元的数据集成方法。用数据元语义树对实体属性概念的内涵进行语义描述,使实体属性的语义能被计算机理解。然后通过语义计算实现精确到实例级的语义映射,从而完成数据集成。

**关键词** 数据集成 数据元 语义树 语义映射

**中图法分类号** TP311.11; **文献标志码** A

在我国信息化建设过程中,由于时间或部门不同,使得各企业开发了众多彼此独立的信息系统,积累了大量丰富的数据资源。随着企业规模的扩大和计算机技术的发展,企业信息化建设势在必行。然而,由于建设时期不同,开发部门不同、使用设备不同、技术发展阶段不同等原因,使得数据存储管理极为分散,造成了大量的数据冗余和数据不一致性,形成了众多的信息孤岛,使得数据资源难以共享访问。数据集成的核心任务是要将互相关联的分布式异构数据源集成到一起,使用户能够以透明的方式访问这些数据源<sup>[1]</sup>。

近年来,基于本体<sup>[2-5]</sup>的数据集成方法研究受到了高度的关注。基于本体的数据集成方法是建立全局本体和具体数据源的局部本体,以及全局本体和局部本体、局部本体之间的映射关系,完成异构数据源之间语义上的映射,最后将这种语义上的对应关系转换为数据查询,完成数据集成。但是基于本体的研究方法仅考虑了模型概念的外延即概念之间的关系,没有考虑概念的内涵,对实体属性的语义只能依靠自然语言描述或分析模型的结构,

缺少对数据模式元素进行精确描述的语义方法。本文提出的基于数据元的数据集成方法,通过对概念的内涵进行形式化的语义描述,可以实现不同模型之间,乃至实例级的语义映射。

## 1 数据元简介

数据元是用一组属性描述定义、标识、表示和允许值的数据单元,是在一定的环境下不必要再细分的最小数据单位。数据元是可识别和可定义的,每个数据元都有其基本属性,如:名称、定义、数据类型、精度、值域等。一个数据元由数据元概念和表示两部分组成。数据元概念(Data Element Concept)是能以数据元的形式表示,且以任何特定的表示法无关的一种概念。当一个表示被联合到一个数据概念时就能够产生一个数据元。

按照国家标准,数据元分为数据元概念、通用数据元、应用数据元。通用数据元提供的是一般的内容,而非具体内容,具体内容则由应用数据元提出。如通用数据元可指“日期”、“姓名”,而引伸出来的应用数据元可以更为具体,如“考试日期”、“入学日期”、“学生姓名”、“教师姓名”等。应用数据元规定为一个独立应用的数据范畴。一个应用数据元必须来自某个通用数据元,遵循通用数据元给出的框架。一个数据模型中的数据项由于限定于数

2011年3月25日收到

第一作者简介:时贵英(1977—),女,汉族,河北石家庄人,硕士,东北石油大学计算机与信息技术学院讲师,研究方向:计算机软件工程与集成技术。E-mail:dqpisgy@163.com。

据模型的应用范围,因此是一个应用数据元。图 1 描述了用于数据元结构的术语与传统的数据建模术语的关联,在数据模型中,一个数据项可以等同于一个数据元,数据模型的数据元名称的典型形式是实体名称和实体属性名称的合成,如图 2 所示。

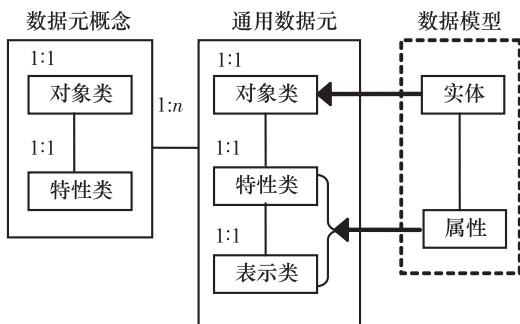


图 1 数据元结构



图 2 数据模型中的数据元

## 2 数据元的语义树<sup>[6]</sup>描述方法

基于数据元的数据集成需要解决的一个关键问题是语义的描述方法,本文采用的是语义树的描述方法,语义树提供了一种形式化的语义描述方式,可以方便地描述数据元的语义。

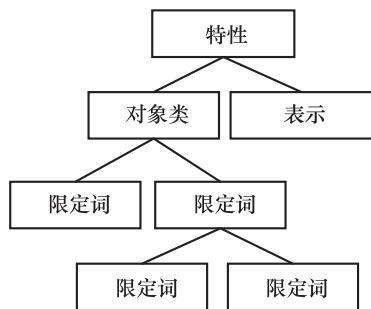


图 3 数据元的语义树

如图 3 所示,在数据元的语义树中,描述数据元“表示”的节点是叶子节点,不再受其它节点限定。

抽象语义树通常对应一个通用数据元,通过附加规则节点可派生各种应用数据元。在数据元的语义树中,特定节点是根节点,对象类可被其它限定词进一步限定。在对数据模型中的数据项进行描述时,由于数据项与具体应用场景有关,所以在相应的数据元的语义描述上增加应用场景约束,才能真正地描述该数据项在数据模型中的语义。数据项数据元以属性、实体、约束为中心,通过对实体进行直接限定和间接限定,构成对数据模型数据项语义的完整描述,图 4 所示。

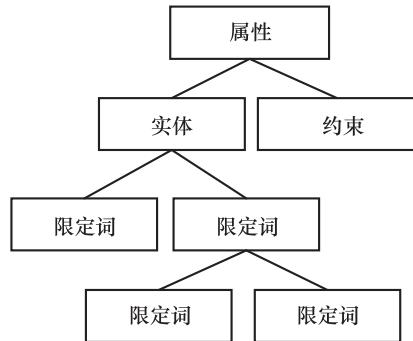


图 4 数据项的语义树

## 3 基于数据元语义树的映射实现

### 3.1 语义树中各节点的权值分配规则

设  $N$  为语义树  $T$  中的任意一个节点,以  $N$  为根的子树记为  $sub(N)$ ,  $N$  的子节点数为  $n$ ,  $N$  的第  $i$  个子节点记为  $child(N, i)$ , 节点  $N$  的权记为  $W(T, N)$ , 子树  $sub(N)$  的权记为  $W(sub(N))$ 。语义树中各节点的权值分配规则如下:

- (1)  $W(T) = 1$ ;
- (2)  $W(sub(N)) = W(T, N) + \sum_{i=1}^n W(sub(child(N, i)))$ ;
- (3)  $W(T, N) = \frac{1}{2}W(sub(N))$ ;
- (4)  $W(sub(child(N, 1))) = W(sub(child(N, 2))) = \dots = W(sub(child(N, n)))$ 。

从上可以看出,一棵语义树的权为 1, 等于语义树中所有节点的权值之和。子树的根节点占子树

权的一半,兄弟节点代表的子树具有相同的权。可见,离语义树的根节点越近权越大,代表的语义概念越重要,这符合一般概念定义的逻辑。

### 3.2 数据元语义树的映射实现举例

假设两个同类数据项  $X$  与数据项  $Y$ ,分属于实体  $t_1$  和  $t_2$ , $X$  和  $Y$  的语义树分别为  $T_x$  和  $T_y$ ,对  $T_x$  和  $T_y$  进行比较,结果如图 5 所示。

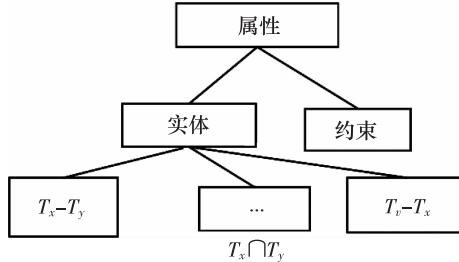


图 5 比较两个同类数据元的语义树

比较结果由  $T_x \cap T_y$ 、 $T_y - T_x$ 、 $T_x - T_y$  三部分组成:

(1)  $T_x \cap T_y$ :数据项相同的部分,是  $X$  和  $Y$  具备同类性的基础。

(2)  $T_y - T_x$ :给出了数据项  $Y$  独特的语义约束部分。如  $Z \in T_y, Z \notin T_x$ ,则  $Z \in T_y - T_x$ 。当  $T_y - T_x$  为空时,表示数据项  $X$  下的实例只是数据项  $Y$  实例的一部分;当  $T_y - T_x$  不为空时,表明数据项  $Y$  下的实例数据只是数据项  $X$  下实例数据的一部分; $T_y - T_x$  用来选择数据实体集。

(3)  $T_x - T_y$ :给出了数据项  $X$  独特的语义约束部分。如  $Z \in T_x, Z \notin T_y$ ,则  $Z \in T_x - T_y$ 。当  $T_x - T_y$  为空时,表示数据项  $Y$  的实例是数据项  $X$  实例的一部分;当  $T_x - T_y$  不为空时,表明数据项  $Y$  下的实例数据只是数据项  $X$  下实例的一部分; $T_x - T_y$  用来决定数据项  $Y$  的实例筛选条件。

例如,学生信息数据库除了存储全体学生的基本信息外,为了研究男女生英语成绩的差异还建立了男生成绩表和女生成绩表。

(1) 表“学生信息”是用来存放学生基本信息的数据表,其每一个实例描述一个学生的基本信息,包括主键学号、姓名、院系等,院系的值是一组枚举值,包括石油工程学院、化学化工学院、计算机

学院、电子科学学院、外国语学院、艺术学院等。

(2) 表“男生英语”用来记录所有院系男生的英语成绩,其中字段学号是外键,引用表“学生信息”中的学号。

(3) 表“女生英语”与表“男生英语”的结构相同,但表中的实例记录的是所有院系女生的英语成绩,其中字段学号是外键,引用表“学生信息”中的学号。

现在,我们需要统计“计算机学院全体学生的英语成绩”,把“计算机学院全体学生的英语成绩”记为  $X$ ,则其对应的语义树  $T_x$  如图 6 所示:

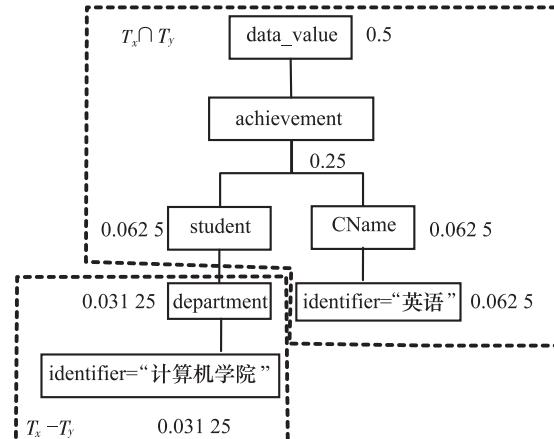


图 6 “计算机学院全体学生英语成绩”的语义树

首先用  $X$  的语义树  $T_x$  与各数据项语义树进行比较,得到两个候选数据项:“男生英语. 成绩”、“女生英语. 成绩”。令  $Y$  为“男生英语. 成绩”,则对应的语义树  $T_y$  如图 7 所示。

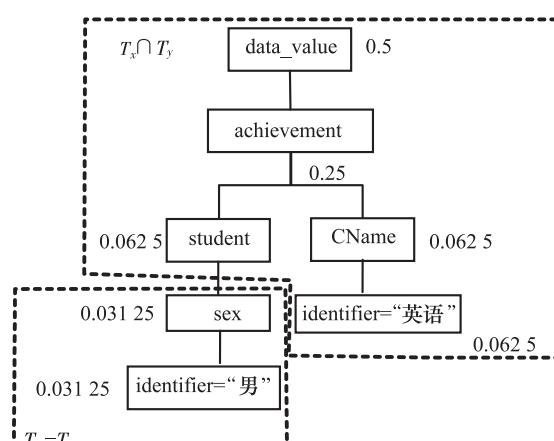


图 7 数据项“男生英语. 成绩”的语义树

比较  $X$ 、 $Y$  的语义树  $T_x$ 、 $T_y$ 。比较结果  $T_x \cap T_y$ 、 $T_x - T_y$ 、 $T_y - T_x$  分别在图 6、图 7 中用虚线框标出。在图 6 中  $T_x - T_y$  中, 可以构成路径 `data_value. achievement. student. department. identifier = "计算机学院"`, 而表“男生英语”的其它数据项的路径都没有与之完全匹配的, 但外键“学号”与之部分匹配, 由于“男生英语. 学号”对应的主键是表“学生信息”中的学号, 因此在表“学生信息”中查找相应的数据项, 得到数据项“院系”的路径与之匹配。因此得到约束条件“学生信息. 院系 = '计算机学院'" 和关联条件“男生英语. 学号 = 学生信息. 学号”, 两者相与就得到筛选条件“男生英语. 学号 = 学生信息. 学号 and 学生信息. 院系 = '计算机学院'"。

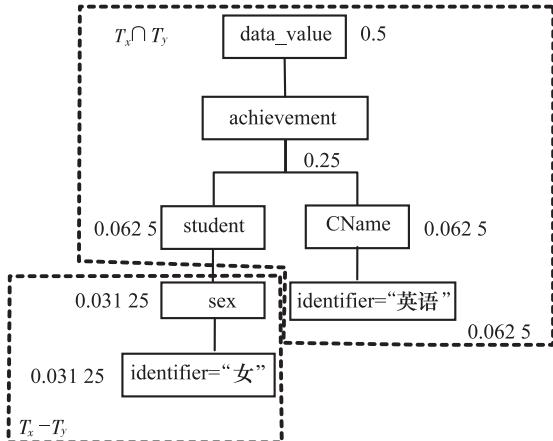


图 8 数据项“女生英语. 成绩”的语义树

同理, 与图 8 中的数据项“女生英语. 成绩”的语义树比较, 得到筛选条件“女生英语. 学号 = 学生信息. 学号 and 学生信息. 院系 = '计算机学院'"。将以上结果用关系数据库 SQL 语句表达, 结果是:

(1) select 男生英语. 成绩 from 男生英语, 学生信息 where 男生英语. 学号 = 学生信息. 学号 and 学生信息. 院系 = '计算机学院';

(2) select 女生英语. 成绩 from 女生英语, 学生信息 where 女生英语. 学号 = 学生信息. 学号 and 学生信息. 院系 = '计算机学院'。

把两个查询结果合并在一起, 可以得到“计算机学院全体学生英语成绩”需要的全部数据。

## 4 结论

由于数据元语义树的子树本身也是一棵树, 因此容易处理子映射。在上节的例子中, “计算机学院全体学生英语成绩”不仅映射到了概念相似的数据项“男生英语. 成绩”和“女生英语. 成绩”, 而且映射到了相关的实体和属性“学生信息. 院系”, 因此相关概念的映射保证了映射概念的语义完整性。在实际项目中, 对 7 000 多条数据元进行了语义描述, 通过映射计算, 均能得到正确的映射结果。

## 参 考 文 献

- 1 Maurizio L. Data integration: a theoretical perspective. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 2002
- 2 周刚, 郭建胜, 石磊. 基于本体的异构数据源集成系统分析与设计. 北京联合大学学报(自然科学版), 2007;21(1):45—48
- 3 张磊, 吴笑凡, 谢强, 等. 基于 Ontology 的多数据源语义集成研究. 信息系统, 2005;28(6):656—659
- 4 Irina P, Heli H, Juha T. Semantic interoperability information integration by using ontology mapping in industrial environment. Proceedings of the 10th International Conference on Enterprise Information Systems, ICEIS 2008, 2008;5: 465—468
- 5 鱼滨, 郑娅峰. 基于本体的异构数据集成方法及其实现. 计算机应用与软件, 2007;24(9):30—33
- 6 Wen Bilong, Zhang Li. Defining semantics for data element with semantic tree. ISISE'2008 (2008 International Symposium on Information Science and Engineering), 2008;524—527

## Research of Data Integration Based on Data Elements

SHI Gui-ying, WEN Bi-long, WANG Zhi-bao

(School of Computer&Information Technology, Northeast Petroleum University, Daqing 163318, P. R. China)

[**Abstract**] For the shortcomings of the attribute-level mapping in the current data integration approach, a new data integration approach based on data elements is presented, which uses semantic tree of data elements to describe the connotation of the properties's concept, and the semantic of the properties can be understood by computer, and then the instance-level semantic mapping can be realized by semantic computation, so that the data integration can be completed.

[**Key words**] data integration    data elements    semantic tree    semantic mapping

(上接第 4218 页)

## An Improved Method of MFCC Parameter Extraction in Speaker Recognition

HE Zhao-xia, PAN Ping

(College of Computer Science & Information, Guizhou University, Guiyang 550023, P. R. China)

[**Abstract**] Speech feature parameter extraction is an very important part of the speech recognition system, especially in speech training and recognition. Mel frequency cepstrum coefficient (MFCC) is a common feature, It can analysis and process speech signal, remove redundant information in speech recognition, and gain important information which influence speech recognition. Owing to time-varying and chaotic characteristic of voice signal, a improved MFCC feature extraction method based on nonlinear stochastic resonance theory is proposed. By comparison results of two methods, it is proved that the improved one is practicable and more superior which provides a new direction of speech feature parameter extraction in speech recognition.

[**Key words**] speech recognition    feature extraction    MFCC    stochastic resonance