专家信息语义模型异构数据转换技术①

杨 佳 黄 芳 龙 军 刘高嵩 (中南大学 信息科学与工程学院 湖南 长沙 410083)

摘 要: 针对专家信息异构数据转换过程中数据之间映射关系的建立问题，提出了一种基于专家信息语义模型的方法。详细阐述了专家数据表语义模型的建立方法，语义相似度计算的方法及基本原理，选取了一种适用于专家数据语义的计算模型，在此模型的基础上对各种影响语义之间相似度的因素进行调整，并对实验结果进行对比分析，找到了一种更加有效、准确的语义相似度计算方法。

关键词: 专家信息; 转换; 语义模型; 异构数据

Heterogeneous Data Conversion Based on Semantic Models of Expert Information

YANG Jia , HUANG Fang, LONG Jun, LIU Gao-Song

(School of Information Science and Engineering, Central South University, Changsha 410083, China)

Abstract: A new method based on semantic models of expert information is proposed. It aims to resolve the problem of establishing data mapping in a heterogeneous data conversion process of expert information. This paper investigates the methods of establishing a semantic model based on a data table of expert information and the basic principles and methods of semantic similarity calculation. It selects a calculation model suitable for semantics of expert data and adjusts the factors which affect the semantic similarity of the data table in this model. Through a comparative analysis of the experimental results, a more effective and accurate method of calculating the semantic similarity is found.

Keywords: expert information; conversion; semantic model; heterogeneous data

1 引言

目前我国各省份以及教育部、国家奖励办等部门都建有自己的专家数据库系统，这些专家数据一般都是异构的，要想让各地的专家信息实现共享，就必须对异构数据进行转换，解决各种数据在数据结构、语义等各方面的冲突[1]。转换后的数据不仅要保持完整性、一致性等特点，还要保持原数据的完整语义，为此我们提出一种基于语义模型的专家信息异构数据转换解决方案。

数据之间映射关系的指定是异构数据转换中的关键问题，传统的异构数据转换方法如基于转换工具、基于中间件数据库、基于 XML 技术等[2-4]，都是通过人工来指定其转换规则和数据之间的映射关系，用基于语义模型的方法可以根据语义之间的相似度来自动产生数据元素之间的映射关系。在当前大多数的专家数据库中，专家的重要信息都分别存放于不同的数据表中，比如专家的基本信息存放于专家表中、熟悉学科信息存放于学科表中、项目信息(主持或参加的基金/计划项目)存放于计划表中等等。这些数据表之间都是通过外键约束联系在一起，具有一定的逻辑关系，当我们想要得到一个专家的完整信息时，就要通过这些逻辑关系从不同的数据表中进行查询、整合。因此在进行专家信息异构数据转换时，就必须先分析出数

① 基金项目:国家自然科学基金(60873081,M0921005,U0835003);高等学校博士学科点专项科研基金(20090162110072)
收稿时间:2010-01-15;收到修改稿时间:2010-02-27

据之间的逻辑关系，进而才能保证转换后专家数据语义的完整性。基于专家信息系统的以上特点，我们可以根据数据表之间的逻辑关系来得到专家信息数据之间的语义模型。

基于逻辑关系的语义模型建立方法是最近几年才提出来的[5]，其特点是适合于分析各种关系模型和嵌套模型的语义关系，例如关系数据库、XML 模型等。并且这种方法也考虑到了语义之间的约束关系，因此能保证源数据和目标数据在语义约束和数据内容上的一致性。这种方法能用于基于源模式和目标模式的数据交换、数据集成，以及基于语义的数据查询和数据重写等方面。在专家数据库中，通常用计算语义之间的相似度来确定专家信息语义之间的匹配关系。语义之间相似度的计算包括基于距离的语义相似度计算模型、基于内容的语义相似度计算模型和基于属性的语义相似度计算模型。对于数据库表而言，表示数据的字段都可以看成是一个属性，因此我们采用基于属性的语义相似度计算模型来确定语义之间的关系，我们用这种方法进行实验，取得了令人满意的结果。

2　专家数据库语义模型设计

专家数据库系统的主要作用是收集相关专家的基本信息和学术信息比如专家所在单位、熟悉学科、评审经历、获奖情况、参加或主持的项目情况等等，并及时更新专家的最新信息，当需要专家参加某项科技评价活动时，确保专家信息的准确性和真实性。专家的这些信息具有一定的逻辑关系，即专家数据信息具有一定的语义联系，在进行数据转换时就必须找到一种能够保持这种语义联系的映射规则。我们通过设计专家数据库的语义模型来得到数据之间的映射规则，分别从专家数据库语义模型的建立过程、专家数据库表关系和具体实现过程三个方面来阐述专家数据库语义模型的设计。

2.1 专家数据库语义模型的建立过程

为了实现数据映射，需要给出一种源数据库和目标数据库语义都一致的方式解释二者之间的对应关系，由于专家信息存放在不同的数据表中，而数据库表之间存在一定的逻辑结构，因此采用通过逻辑关系来得到数据之间的语义。专家数据库语义模型的建立过程如下：

1) 得到数据表的原始语义。根据需要进行转换的专家数据信息，确定将要进行转换的数据库表及表中的字段，得到每张数据表的原始语义。

2) 将约束关系转化为语义。专家数据表之间存在很多约束，如外键约束、学科代码的唯一性约束等，将所有涉及到的约束转化为语义。

3) 在数据表原始语义的基础上加入约束关系，得到数据表完整语义。将每张数据表所涉及的约束关系全部加入进来，得到此数据表的完整语义信息。

2.2 专家数据表关系

在专家数据库中，专家的信息都分散存放在很多数据表中，为了说明专家数据库语义模型的建立过程，我们选取包含专家大部分重要信息的三张数据表，单位表(T_DWUNIT)、专家表(T_ZJSPRCIALIST)和熟悉学科表(T_ZJSPECIALIST_XK)，三个数据表之间的关系如图 1 所示：

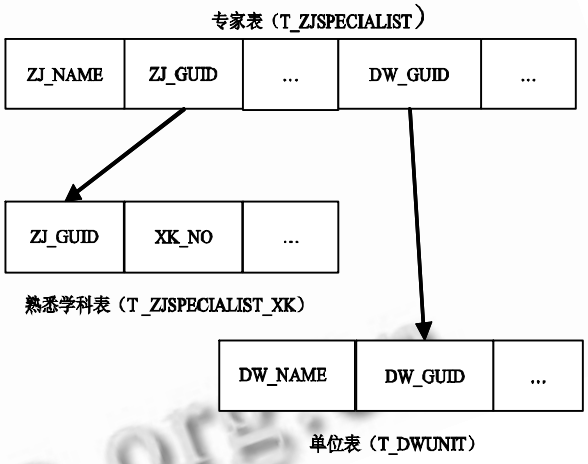

图 1　专家表、单位表和熟悉学科表之间的关系

专家的基本信息存放在专家表中，专家的熟悉学科信息存放在熟悉学科表中，专家的所在单位信息存放在单位表中。专家表中的 ZJ_GUID 是专家的唯一标识，不可重复的 32 位 GUID 号，其对应着学科表中的 ZJ_GUID，如果学科表中的 ZJ_GUID 与专家表中的 ZJ_GUID 相同，则这条熟悉学科信息就属于这个专家。同样，专家表中的 DW_GUID 也是单位的唯一标识，对应着单位表的 DW_GUID，当查看专家所在单位信息时就要查询单位表中与专家表 DW_GUID 相同的那条记录。

2.3 实现过程

为了便于数据表中的数据在语义中的表示，我们使用如下的子句。如果要指定数据表的某一列，用 x in TableName 表示；如要指定某列所代表的值，用

x.TableName 表示。用 Select *表示数据表中的所有元素。语义模型的建立分两个步骤进行：

1) 通过数据表得到原始的语义。

此时只考虑数据表的属性，不考虑数据表之间的约束关系，对于图 1 中三个数据表得到如下的语义：

S1: Select * from d in T_DWUNIT

S2: Select * from z in T_ZJSPECIALIST

S3: Select * from x in T_XK_SPECIALIST

2) 数据表之间的约束关系

将数据表之间的约束以$\forall P1 \exists P2 B$的形式给出，其中$P1$、$P2$分别为数据库表所代表的语义，B为相等的条件，对图 1 中的外键约束可以描述为：

(a) for z in T_ZJSPECIALIST
exists x in T_XK_SPECIALIST
where z. ZJ_GUID=x. ZJ_GUID

(b) for z in T_ZJSPECIALIST
exists d in T_XK_SPECIALIST
where z..DW_GUID=d. DW_GUID

3) 加入数据表之间的约束关系得到完整语义

在得到原始语义和约束关系后，采用 Chase 方法追踪所有的外键约束，产生逻辑相关的最大数据元素集合，得到数据表完整的语义，图 1 中三个数据表的完整语义为：

S1´:select d.DW_GUID,d.DW_NAME, … from d in T_DWUNIT

S2´: select z.ZJ_GUID,z.ZJ_NAME,z.DW_GUID, …, x.K_NO,…,d.DW_NAME,…

from z in T_ZJSPECIALIST, x in T_XK_SPECIALIST, d in T_DWUNIT

S3´:select x.ZJ_GUID, x.XK_NO ,… from x in T_XK_SPECIALIST

3 语义相似度计算

在得到专家信息数据的语义之后，为了确定这些语义之间的相互联系，就要对语义进行匹配。在专家数据库中，我们采用计算语义相似度的方法来匹配语义之间的联系。

3.1 语义相似度计算模型

传统的语义之间相似度的计算模型有三种：基于距离的语义相似度计算模型、基于信息容量的语义相似度计算模型和基于属性的语义相似度计算模型。基于距离的语义相似度把概念之间的关系表示成一个有向图，用两个概念在有向图中的几何距离来衡量语义相关度，距离越短则相关度越高；基于信息容量的算法[6]的基本原理是："如果两个概念共享的信息越多，则它们之间的语义相似度就越大，反之越少，就越小。"基于属性的相似度算法是根据比较两个概念间相同的和不同的属性多少来确定二者的相似度。Tversky 在他的文章中提到了这种基于属性的计算模型[7]，假设存在两个概念和，那么他们的相似度计算公式：

$$Sim(s1,s2)=q\,f(s1\cap s2)-\partial f(s1-s2)-b\,f(s2-s1) \quad (1)$$

在式（1）中$q,\partial,b\geq 0$；$(s1\cap s2)$为$s1$和$s2$的共有属性；$s1-s2$表示$s1$有而$s2$没有的属性；$s2-s1$表示$s2$有而$s1$没有的属性。参数$q,\partial,b\geq 0$是为概念 s1 和 s2 之间的相同属性和不同属性而赋予的不同权重。之所以要对不同的属性赋予不同的权重，是因为考虑到相同属性和不同属性在区别不同的概念的过程中所起的作用不一样。

3.2 专家属性语义相似度计算策略

在专家数据库中，专家的信息都存储在数据表中，而数据表中的字段信息都可以看成是属性集合，我们得到的语义信息就可以看成是很多属性组成的集合。在不同的专家库系统中，表示专家重要信息的字段名称基本上是相同的，如专家编号的字段 ZJ_NO，专家姓名的字段 ZJ_NAME，熟悉学科代码 XK_NO 等，只是这些信息所存放的数据表不同，因此在专家数据的语义中，很多语义所包含的属性名称是相同的，所表示的数据信息也是能够匹配的。针对这些特点，专家信息数据语义之间的关系就可以通过计算语义中属性的匹配程度来决定。

3.2.1 语义属性集合

根据基于属性的语义相似度计算模型，我们将每张数据表所代表的语义关系中相关的元素作为属性放入到一个集合中，图 1 中三张表所对应的语义属性集合分别为 S1、S2、和 S3。

S1 包含单位信息，其属性集合为：DW_GUID, DW_ID,DW_NAME,DW_ADDRESS,DW_POSTCODE,DW_FAX,DW_HIERARCHICAL_SN,DW_PARENT_GUID,DW_T

YPE,DW_EMAIL,DW_UNPROCODE.

S2 包含专家基本信息、单位信息和熟悉学科信息，其中：

专家基本信息属性集合为：ZJ_GUID,ZJ_NO,ZJ_NAME,XB_SEXCODE,ZJ_ID,WP_EDUCODE,XW_DEGCODE,DW_GUID,DW_HIERARCHICAL_SN,ZJ_MPNO,ZJ_STATE,ZJ_SUBMITTED,ZJ_EMAILBOX,ZJ_BIRTHDAY,ZJ_IDTYPE,DP_SECTCODE,MZ_NATIONCODE,SF_BODAO,YS_ACADECODE;

单位信息属性集合为：DW_ID,DW_NAME,DW_ADDRESS,DW_POSTCODE,DW_FAX,DW_HIERARCHICAL_SN,DW_PARENT_GUID,DW_TYPE,DW_EMAIL,DW_UNPROCODE;

熟悉学科信息属性集合为：XK_REMARK,XK_NO,XK_GUID, SX_DEGREE;

S3 包含熟悉学科信息，其属性集合为：ZJ_GUID,XK_REMARK,XK_NO,XK_GUID, SX_DEGREE.

我们从源数据库中选取的单位表、专家表和学科表三张表，对应的语义为 S1、S2 和 S3；从目标数据库中选取单位表和专家表两张表，对应的语义为 T1 和 T2。T1 和 T2 所对应的语义属性集合分别为：

T1 包含单位信息，其属性集合为：DW_GUID,DW_NAME,DW_ID,DW_HIERARCHICAL_SN,DW_TYPE,DW_PARENT_GUID,DW_ADDRESS,DW_POSTCODE,DW_FAX,DW_EMAIL,JL_CHARGE_DEPT,SHARE_PLAN;

T2 包含专家基本信息和学科信息，其中熟悉学科代码作为基本信息的一部分，其属性集合为：XM_ID,ZJ_NO,ZJ_NAME,ZJ_ID,BM_RECOMDEPT,XK_NO,SF_BODAO,YS_ACADECODE,DW_GUID,DW_NAME,DW_HIERARCHICAL_SN,XB_SEXCODE,ZC_TITLECODE,WP_EDUCODE,XW_DEGCODE,ZJ_EMAILBOX,ZJ_MPNO,ZJ_BIRTHDAY,ZJ_IDTYPE,DP_SECTCODE,MZ_NATIONCODE;

3.2.2 相似度计算策略

对于公式(1)中的函数 f，我们将其定义为计算相同或不同属性之间的个数，从以下几个方面进行了实验：

1) 只考虑语义之间的相同属性

专家数据转换的目的是要从源数据库中得到需要的相关数据，因此在进行语义匹配时只要目标语义中的属性在源语义中存在即可。

此时我们设置 $q=1,\partial=0,b=0$，此时语义之间的相似度值可以表示为：

$$Sim(s1,s2)=\frac{f(s1\cap s2)}{f(s1)} \tag{2}$$

由于重点考虑目标语义中属性能够匹配的程度，因此只计算相同属性在目标语义中所站的比重，匹配后的结果如表 1 所示。

表 1　只考虑相同属性的语义匹配结果

目标语义	源语义	匹配结果
T1	S1	83.33%
T1	S2	91.67%
T1	S3	0%
T2	S1	14.29%
T2	S2	90.48%
T2	S3	4.67%

根据上面的结果，目标语义 T1 与源语义 S2 的匹配程度最高，而在实际应用中，T1 应该与 S1 最为匹配，这是因为源语义 S2 包含了语义 S1 中的属性元素，因此只考虑概念间相同属性的方法并不适合。

2) 考虑语义之间的不同属性

由于语义中存在很多不同的属性，有些时候两条语义中不同属性的数量可能比相同属性的数量还多，这种情况下只考虑相同属性的匹配结果是不正确的。此时我们考虑目标语义中在源语义中不存在的属性对语义相似度的影响，设置，此时语义间的相似度值可以表示为：

$$Sim(s1,s2)=\frac{f(s1\cap s2)-f(s1-s2)}{f(s1)+f(s2)-f(s1\cap s2)} \tag{3}$$

此时还要考虑到源语义中不存在于目标语义中的属性集对结果的影响，因此要计算源语义与目标语义的整个属性集，匹配后的结果如表 2 所示：

表 2　考虑不同属性影响的匹配结果

目标语义	源语义	匹配结果
T1	S1	61.54%
T1	S2	29.41%
T1	S3	-70.59%
T2	S1	-51.72%
T2	S2	48.57%
T2	S3	-76.00%

在上面的结果中，出现负数是因为语义之间不同的属性比较多。这种方法能比较直观、准确的反应出语义之间属性的匹配程度。

3) 考虑关键属性的权值

在专家信息系统中，有些专家的信息是非常重要且必不可少的，因此在语义匹配时要按属性的重要性来赋予不同的权值。我们将目标语义中专家的重要基本信息如专家编号、专家姓名、熟悉学科代码、推荐单位等，以及专家的重要单位信息如单位 GUID 号、单位编号、单位层次号、单位类型等作为关键信息，并给这些信息赋予权值 2。

此时我们将目标语义分成两部分：关键信息和非关键信息。目标语义 T1 分为 T11(关键信息)和 T22(非关键信息)，T2 分为 T21(关键信息)和 T22(非关键信息)，分开后的属性集如下：

T11(DW_GUID,DW_NAME,DW_ID,DW_HIERARCHICAL_SN,DW_TYPE,DW_PARENT_GUID)

T12(DW_ADDRESS,DW_POSTCODE,DW_FAX,DW_EMAIL,JL_CHARGE_DEPT,SHARE_PLAN)

T21(XM_ID,ZJ_NO,ZJ_NAME,ZJ_ID,BM_RECOMDEPT,XK_NO,SF_BODAO,YS_ACADECODE,DW_GUID,DW_NAME,DW_HIERARCHICAL_SN)

T22(XB_SEXCODE,ZC_TITLECODE,WP_EDUCODE,XW_DEGCODE,ZJ_EMAILBOX,ZJ_MPNO,ZJ_BIRTHDAY,ZJ_IDTYPE,DP_SECTCODE,MZ_NATIONCODE)

设置$q=2,\partial=1,b=0$，此时语义间的相似度可以表示为：

$$Sim(s1,s2)=\frac{q f(s11\cap s2)+\partial f(s12\cap s2)-q f(s11-s2)-\partial f(s12-s2)}{q f(s11)+\partial f(s22)+f(s2)-f(s1\cap s2)} \quad (4)$$

式(4)中的 S11 表示目标语义中的关键信息，S12 表示目标语义中的非关键信息，匹配后的结果如表 3 所示：

表 3　考虑关键属性权值的匹配结果

目标语义	源语义	匹配结果
T1	S1	73.68%
T1	S2	45.00%
T1	S3	-78.26%
T2	S1	-50.00%
T2	S2	56.52%
T2	S3	-77.78%

对比表二中的结果可知，这种方法在保证准确性的前提下，得到的匹配结果更令人满意。

4　根据语义匹配度得到映射关系

在数据转换中，关键的问题就是要得到数据间的映射关系，我们可以根据计算出的语义之间相似度的匹配结果来建立映射关系。

根据语义相似度的计算结果，我们得到了两组互相匹配的语义，即目标语义 T1 与源语义 S1 匹配，目标语义 T2 与源语义 S2 匹配。因此，目标语义 T1 的数据元素可以由源语义 S1 中的数据元素映射得到，目标语义 T2 的数据元素可以由源语义 S2 中的数据元素映射得到。由于语义中的数据元素来自不同的数据表，通过分析相互匹配的数据元素在各自的专家数据库中分别来自哪张数据表，我们可以得到源专家数据库与目标专家数据库数据表之间的映射关系，我们根据选取的实例所表示的语义模型，得到的数据映射结果如图 2 所示。

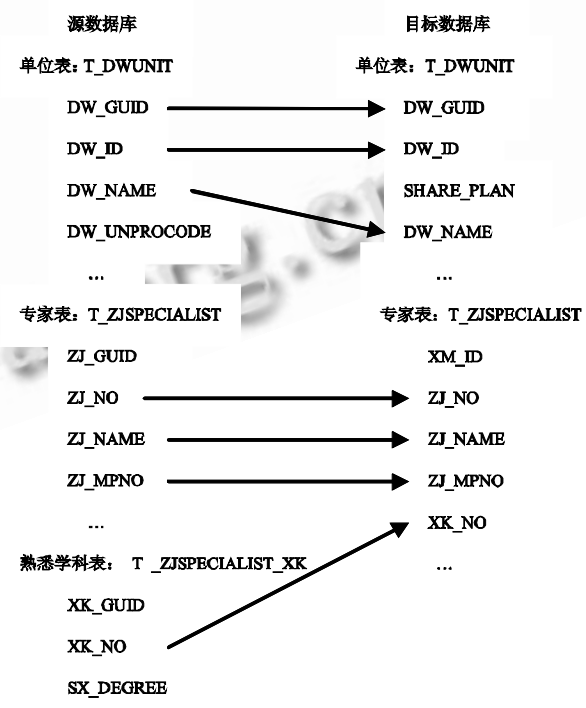

图 2　专家数据表映射关系

在图 2 中箭头即表示数据元素之间的映射关系。目标语义 T1 中的单位信息如 DW_GUID、DW_ID 和 DW_NAME 来自目标数据库的单位表，这三个字段又

是从源语义 S1 中映射过来，而 S1 中的这三个字段来自于源数据库的单位表，由此建立了源数据库和目标数据库单位表之间的映射关系。同样，目标语义 T2 中的专家基本信息如 ZJ_NO、ZJ_NAME 和 ZJ_MPNO 可以由源数据库的专家表映射得到；目标语义 T2 中的 XK_NO 对应源语义 T1 中的 XK_NO，而 T1 中的 XK_NO 来自于源数据库的熟悉学科表，因此得到了源数据库熟悉学科表与目标数据库专家表的映射关系。在建立映射关系时，目标数据库中的有些字段由于在源数据库中匹配不到而缺少映射关系，如图 2 中所示的 SHARE_PLAN 和 XM_ID 两个字段。由此我们根据语义相似度的计算结果自动产生了数据元素之间的映射，从而为专家信息异构数据的转换提供了良好的条件基础。

5 结论

本文的主要工作是建立了专家数据表之间的语义模型，找到了一种计算语义之间的相似度方法来自动生成数据之间的映射关系。在语义相似度计算方面分别从值考虑相同属性对语义关系的影响、同时考虑相同属性和不同属性对语义关系的影响以及考虑关键属性的权值对语义关系的影响三个方面进行了研究，通过实验表明，考虑关键属性的权值对语义关系的影响这种方法能更好的表达出语义之间的匹配关系，为数据元素映射关系的产生提供正确的条件基础。语义相似度的计算是是整个数据转换的基础，本文只是对计算方法进行了初步的研究，将来可以找到一种方法用来更精确的计算语义之间的匹配程度，为数据转换提供更可靠的支持。

参考文献

1 唐红军,万健.利用数据库转换实现异构数据库数据共享的研究与实现.计算机与数字工程, 2005,33(6):99－102.

2 蔡延峰,蔡启明.异构数据库间的数据转换. 计算机与现代化, 2002, (1);42－43,50.

3 谢莉莉,林春梅,陈家训.基于 XML 的数据交换中心模型研究. 东华大学学报(自然科学版), 2001,27(6):33－36.

4 江兵,沈叶忠.高校内部 MIS 的集成与信息交换. 河海大学常州分校学报, 2003, (3):29－33.

5 Popa L, Velegrakis Y, Miller RJ, Hernandez MA, Fagin R. Translating web data. Proc. of the Internet. Conference on Very Large Data Bases (VLDB), 2002. 598－609.

6 Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proceeding of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995. SanFrancisco: Morgan Kaufmann Publishers, 1995:448－453.

7 Tversky A. Features of Similarity. Psychological Review, 1977,84(4):327.