第7卷第2期　　空 军 工 程 大 学 学 报(自然科学版)　　Vol.7 No.2
2006年4月　　JOURNAL OF AIR FORCE ENGINEERING UNIVERSITY(NATURAL SCIENCE EDITION)　　Apr. 2006

基于限制的 OLAM 任务语义模型

段　艳，　程英蕾
(空军工程大学 电讯工程学院，陕西 西安　710077)

摘　要：通过对大量决策分析算法的分析，提出了基于限制(Constraint)的 OLAM 任务语义模型，该模型从一定的逻辑高度对 OLAM 进行抽象，利用它可以分析 OLAM 任务的共性，提取 OLAM 任务执行时的基本操作，这些基本操作的基础是语义片断。文中给出了语义片断之间关联关系的分析和判断方法，提出了基于限制的、具有方向性的语义关联度定义。语义片断间关系的确定是 OLAM 任务/事务优化的基础。

关键词：联机分析挖掘；任务语义模型；语义片断
中图分类号：TP311　　**文献标识码**：A　　**文章编号**：1009－3516(2006)02－0090－05

联机分析挖掘(OLAP－Mining，OLAM)是将联机分析处理(Online Analytical Processing，OLAP)技术和数据挖掘(Data Mining，DM)技术相结合而建立的新的交互式分析技术，一方面 OLAP 的分析结果给 DM 提供挖掘的依据，引导 DM 的进行；另一方面，在数据挖掘的结果中进行 OLAP 分析，则 OLAP 分析的深度就可拓展。但它不是两者单纯的叠加。OLAM 将更能适应实际需要，为单位管理和决策活动提供了一个新的工具，也为决策支持系统的研制提供了新思路。

本文研究的目的就是解决 OLAM 系统中 OLAM 语言的设计和语法、语义分析与优化的问题，为 OLAM 系统的进一步研究和应用提供理论基础和设计参考。

1　OLAM 任务语义模型

1.1　任务抽象模型

在对 OLAM 任务进行深入分析时，需要在一定逻辑高度上对 OLAM 任务进行抽象，建立 OLAM 任务抽象模型。研究人员在对数据挖掘技术进行研究时，为了说明数据挖掘的过程提出了一种简单的表述[1]，模式是一个用语言 L 来表示的表达式 E，它可用来描述数据集 F 中数据的特征，E 所描述的数据是集合 F 的一个子集 F_E，E 作为一个模式要比列举数据子集 F_E 中所有元素的描述方法简单。这种对数据挖掘的简单表示由于抽象程度太高，而在具体数据挖掘算法的研究中没有发挥太大的作用。

我们在 OLAM 的研究中通过对大量数据决策分析算法(数据分析、数据挖掘算法)的深入分析提出了一种新的、实用的 OLAM 任务抽象模型——OLAM 任务语义模型。OLAM 任务的执行过程就是在“人”的参与下，利用领域知识对已知句子集进行充分理解，从中发现满足目标语义的句子的过程。它可以表示为 $T=AD(D,K,U)$

OLAM 任务是在分析人员意图 U 的引导下，利用已知的领域知识 K，对表示原始粗糙语义的句子集(原始数据) D 进行分析求精，对表示目标语义的句子(集) T 的语义进行逐步丰富、完善的过程。这种关系可以用图 1 来表示在这个过程中需要利用领域知识来建立语义环境的语法和语义规则，OLAM 任务正是在这些语法和语义规则的约束下，对语义进行逐步求精的过程。

抽象模型的具体工作过程如图 2 所示，主要包括以下几个基本操作：

收稿日期：2005－07－11
基金项目：军队科研基金资助项目
作者简介：段　艳(1982－)，女，陕西西安人，硕士生，主要从事计算机网络与数据库技术研究.

1）兴趣语义片断的外延扩展

设兴趣语义片断为 P,则语义片断外延的扩展可以表示为 $P'=M(P,D)$,其中 $M(\cdot)$ 为语义片断外延扩展函数,根据兴趣语义片断 P 在语义粗糙的原始句子集 D 中获得与兴趣语义片断相关的句子或语义片断 P'。

2）语义内涵的分析

表示为 $P''=A(P')$,其中 $A(\cdot)$ 负责对 P' 进行内涵分析,生成一个或多个语义片断和句子 P''。

3）语义片断的理解

表示为 $P^*=E(P'')$,其中 $E(\cdot)$ 负责从多个语义片断 P'' 中决定一组分析人员感兴趣的(或与多目标语义相似度较高)的语义片断,我们称之为“兴趣语义片断”,P^* 是兴趣语义片断集。

4）目标语义的完善

表示为 $T^+=S(P^*)$ 或 $T=T+S(P^*)$,$S(\cdot)$ 负责在兴趣语义片断集 P^* 中按某种标准确定属于目标语义的语义片断,并用这些语义片断来丰富目标语义。

从上面的分析来看,OLAM任务模型中一个主要内容是语义片断的表示,语义片断的表示能力直接影响到OLAM任务抽象模型的表达能力和覆盖范围。初步分析会发现:不同的OLAM任务类型其所对应的语义片断的表示形式也各不相同。这成为OLAM任务抽象模型应用的一个主要障碍。

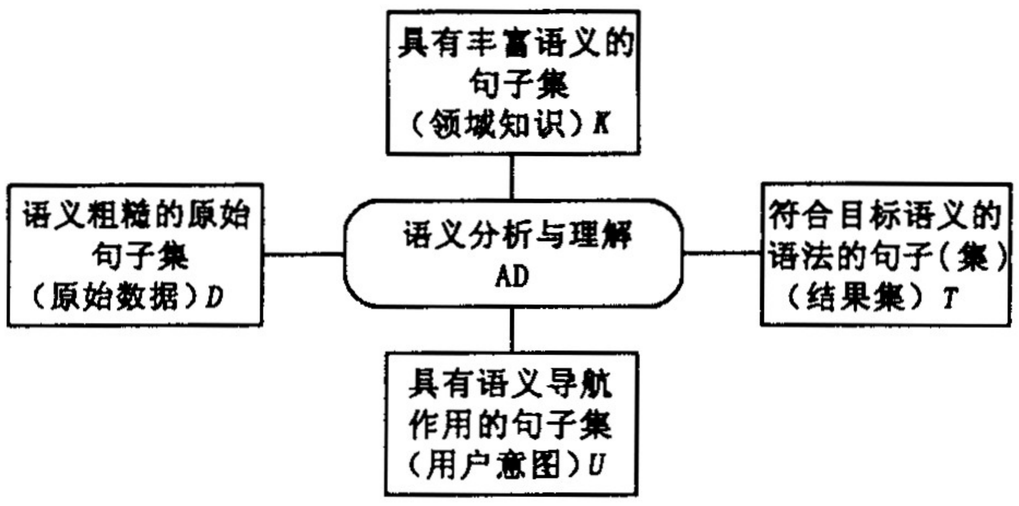

图1　OLAM任务抽象模型结构图

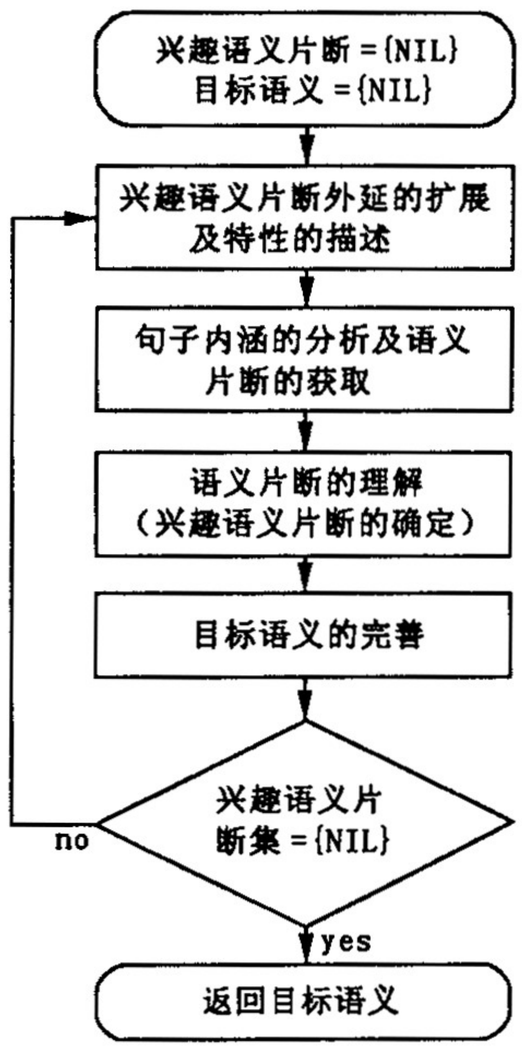

图2　OLAM任务抽象模型工作流程图

1.2　语义片断

为了能够形成统一的语义片断模型,我们引入了“限制数据库”研究领域中的“限制(约束)”的概念。限制是信息的一种表示形式,一定的限制集表示一定的语义的知识,因此用限制来表示信息处理对象是可行的[2-4]。

1.2.1　语义片断的定义

已知对象集 O,$O.P$ 表示对象 O 的特征集,$P_i \in O.P$ 表示对象集的一个特征($i=1,\cdots,n$),$D(P_i)$ 为特性 P_i 的值域。V_{ij} 为 P_i 的值域 $D(P_i)$ 中的一个取值,有 $V_{ij} \in D(P_i)$,则对象 O 上的一个限制 C_1 表示对对象某一个特性的限制(约束)。

即 C_i 为一个析取式:$C_i = P_i\theta V_{i1} \vee \cdots \vee P_i\theta V_{ij} \vee \cdots \vee P_i\theta V_{in}$,也可以表示对象一组特性的线性约束和多项式约束:$C_i=(a_1P_1 + a_2P_2 + \cdots + a_kP_k)\theta a_{k+1}$ 或 $C_i = f(P_1,P_2,\cdots,P_k)\theta$,其中,$\theta=(>,\geqslant,=,\leqslant,<,\neq)\cup(\in,\notin,\subset,\not\subset)\cup(\text{in, between, include})$。

O 上的所有限制(约束)的集合构成 O 上的语义空间 Γ。

对象集 O 上的一个句子 S 为对象集 O 上每个特性的约束的合取式:$S = C_1 \wedge \cdots \wedge C_N$。

句子 S 全面地描述了 O 中部分对象的特性,我们用 $S(O)$ 表示这部分对象(满足 S 的对象集合),则 S 表达了 $S(O)$ 的完整语义。

一个句子中的某一组相关对象的特征 P' 上的限制(约束)称为自由约束,当且仅当 $P_j \in P^f$,P_j 可以取 $D(P_j)$ 中的任意值。此时在表示一个句子时可以忽略这些特征,简略的表示为 $S=\wedge_j C_i(P_j \in P$ 但 $P_j \notin P^f)$,即句子中没有显示表达的对象特性上的限制为自由约束。

对象集 O 上的对某组特性 $P^0 \in O.P$ 上的限制的合取式 C 为一个子句,表达了部分语义,成为语义片断:$C=C_1 \wedge \cdots \wedge C_n = \wedge_j C_j(P_j \in P^o)$。

用 $C(O)$ 表示符合这些语义片断的对象集,有时也把 $C(O)$ 称为语义片断 C 的语义覆盖范围。

句子和语义片断的不同:句子 S 中的每个对象的、每个特性的约束均已确定,表达了对象集中 $S(O)$ 的全部语义。而语义片断 C 中部分对象的部分特性上的约束尚未确定,虽然 $C(O)$ 确定了对象集中的部分对象,但这些对象的语义并不完整,因为这些对象中的部分语义尚未确定。当已知 $S=\wedge_j C_j(P_j \in P)$、$C=\wedge_j C_j(P_j \in P^0)$ 且有 $P^0 \in P$,则有 $S(O) \in C(O)$ 成立。

1.2.2 任务抽象模型与具体算法的映射

OLAM 语义抽象模型具体具有较宽的覆盖范围,可以表示常规的 OLAM 任务,图 3 给出模型与常见的数据挖掘算法的对应关系。

下面以简单关联规则的发现过程来验证 OLAM 任务抽象模型的有效性。

1) 设目标语义片断 $T=\{\text{NIL}\}$;兴趣语义片断集 $C=\{\text{NIL}\}$; 交易记录集为 D ;

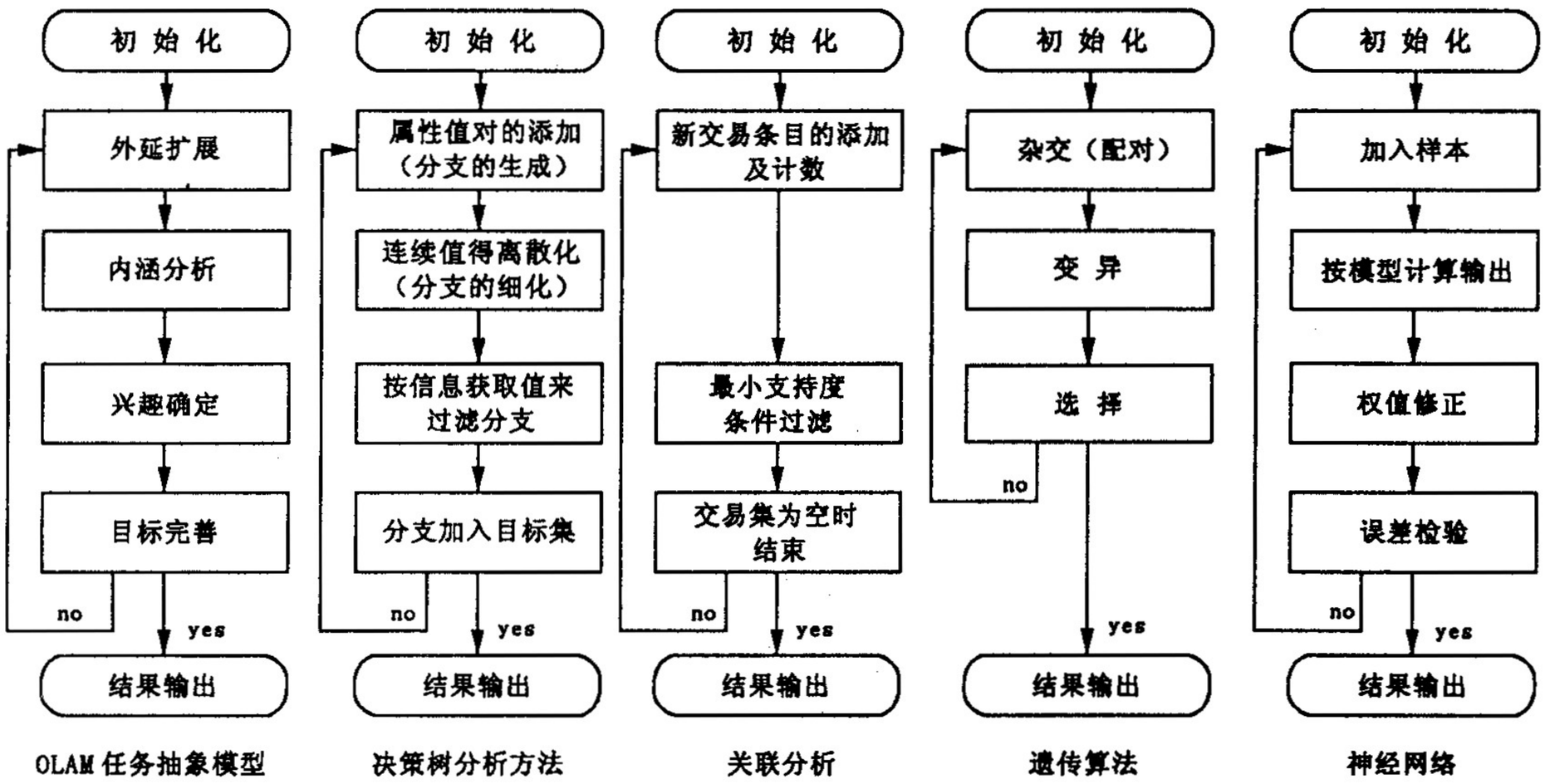

图 3 任务抽象模型与具体算法对照图

2) 语义外延的扩展:根据现有语义片断 C 从交易记录(一组句子)D 中取出一组语义片断来对现有语义片断进行扩展,则 $C=\{\text{NIL}\}\vee\{\text{Transaction}=i\}=\{\text{Transaction}=i\}$ 为扩展后的结果。

3) 语义内涵的分析:按模式对象的模式结构限制的要求对语义片断进行分割,则

$$C=\{\text{Transaction}=i\}\rightarrow C=\begin{cases}\{\text{Transaction}=i\}\wedge\{\text{body}=b_1\}\wedge\{\text{head}=h_1\}\wedge\{\text{Tran. count}>0\}\\ \vdots\\ \{\text{Transaction}=i\}\wedge\{\text{body}=b_n\}\wedge\{\text{head}=h_n\}\wedge\{\text{Tran. count}\leqslant 0\}\end{cases}$$

同时去除 $\text{body}\wedge\text{head}\neq ?$ 的语义片断。

4) 兴趣语义片断的确定:

$$\text{计算 } C=\begin{cases}\{\text{body}=b_1\}\wedge\{\text{head}=h_1\}\wedge\{\text{conf}=c_1\}\wedge\{\text{sup}=s_1\}\\ \vdots\\ \{\text{body}=b_n\}\wedge\{\text{head}=h_n\}\wedge\{\text{conf}=c_n\}\wedge\{\text{sup}=s_n\}\end{cases}$$

5) 当交易记录 D 非空时,则转(2)。当交易记录 D 为空时,利用模式对象的支持度和可信度的约束来对语义片断进行过滤(语义收缩操作),即:

$$C=\begin{cases}\{\text{body}=b_1\}\wedge\{\text{head}=h_1\}\wedge\{\text{conf}=c_1\}\wedge\{\text{sup}=s_1\}\}\wedge\{\text{conf}=c_{\text{val}}\}\wedge\{\text{sup}=s_{\text{val}}\}\wedge\{\text{count}\neq 0\}\\ \vdots\\ \{\text{body}=b_n\}\wedge\{\text{head}=h_n\}\wedge\{\text{conf}=c_n\}\wedge\{\text{sup}=s_n\}\}\wedge\{\text{conf}=c_{\text{val}}\}\wedge\{\text{sup}=s_{\text{val}}\}\wedge\{\text{count}\neq 0\}\end{cases}$$

同时去掉不属于语义空间内有效语义的语义片断,即去掉语义矛盾的语义片断。

6) 此时语义片断所覆盖的范围为目标语义所能表达的语义范围,目标语义中的语义片断为最终的规则。记 $T:=C$。

同理,可以把其它 OLAM 决策分析任务与抽象模型进行对应。

2 语义片断语义相关分析

2.1 语义关联度定义

由于 OLAM 任务抽象模型所描述的 OLAM 任务执行过程是一个语义迭代求精的过程,且由于处理的数

据量较大,每次关于语义片断的语义覆盖范围的确定均从头计算是需消耗很多的资源;另外,每次迭代所涉及的语义片断均有一定的相关性,如果能充分利用这种相关性将会减少计算量,从而提高算法的计算性能。

下面就对语义片断的语义相关性进行讨论。

已知两语义片断 C' 和 C'' 属于同一个语义空间 Γ ,则 C' 与 C'' 之间的语义关联度 $R(C',C'')$ 定义为

$$R(C',C'') = \sum_{i}^{m} w_i r(v'_i,v''_i) \quad (i = 1,\cdots,m)$$

其中 m 为 C' 与 C'' 中相同特性的个数;v'_i 与 v''_i 分别为特征 P_i 在语义片断 C' 与 C'' 上的限制值;w_i 为 P_i 的权值;$r(v'_i,v''_i)$ 为同一对象的属性 P_i 上的语义关联度。$r(v'_i,v''_i)$ 的具体定义与 P_i 所处的语义环境以及 v'_i 与 v''_i 的数据类型有关,一般可以把 v'_i 与 v''_i 的数据类型分为两类:离散值集合和连续值集合。

1) 离散值上 $r(v'_i,v''_i)$ 的定义

设 $v'_i = \{a_1,a_2,\cdots,a_n\}$、$v''_i = \{b_1,b_2,\cdots,b_m\}$ $(m \geqslant n)$,分别为两个有限集,则 $r(v'_i,v''_i)$ 的定义可以借鉴一般有限集间的语义关联度的定义[5],有:

$$r(v'_i,v''_i) = \max_{i \leqslant j \leqslant P_m^n} \left(\frac{1}{n}\sum_{i=1}^{n} (r(a_i,b_{ij}))^p\right)^{\frac{1}{p}}$$

这种语义关联度的定义是在已知集合中每个元素 a 与 b 之间的语义关联度 $r(a,b)$ 以后,利用明可夫斯基距离来定义两有限集之间的语义关联度。

当 $0 \leqslant r(a,b) \leqslant 1$,则 $0 \leqslant r(v'_i,v''_i) \leqslant 1$;由于 $m \geqslant n$,当 $m \neq n$ 时, $r(v'_i,v''_i) \neq r(v''_i,v'_i)$。

2) 连续值集合上 $r(v'_i,v''_i)$ 的定义

把上述定义在离散值集合基础上的语义片断的语义关联度扩展到连续值集合上。该定义同时适用于闭区间和半开半闭区间。

设 $v'_i = (a_1,a_2)$、$v''_i = (b_1,b_2)$ 分别表示两个连续值区间。已知 v'_i 中的任一元素 a 和 v''_i 中的任一元素 b 之间的语义关联度 $r(a,b)$,有 $0 \leqslant r(a,b) \leqslant 1$,则 $r(v'_i,v''_i)$ 定义如下:

$$r(v'_i,v''_i) = \max_{(b_i,b_j) \subset (b_1,b_2)} \left(\frac{1}{r(a_1,a_2)r(b_i,b_j)}\int_1^2\int_i^j (r(a_x,b_y))^p \mathrm{d}a_x\mathrm{d}b_y\right)^{\frac{1}{p}}$$

同样有如下关系 :$0 \leqslant r(v'_i,v''_i) \leqslant 1$;$r(v'_i,v''_i) \neq r(v''_i,v'_i)$。

2.2 $r(v'_i,v''_i)$ 统一简化定义形式

上述语义关联度的定义方法,存在着如下的问题:

1) 计算复杂度高:不但包含幂运算和多重积分,而且每个计算均是一个寻优的过程;在对OLAM任务进行优化时需要大量的语义关联度计算,复杂度较高的语义关联度计算方法将对 OLAM 系统的性能产生不利的影响。

2) 语义不对称:由于 $r(v'_i,v''_i) \neq r(v''_i,v'_i)$,使得对2个语义片断进行语义关联度计算时要计算2次,使得计算效率下降。

3) 定义不统一:离散值与连续值上的语义关联度的定义方式不同,具体实现时的计算方法也不同,给语义片断的计算带来不便。

实际上由于语义片断的语义具有模糊性,语义片断间的语义关联度应该只是一个定性的值,而不一定是定量值。因此可以在降低计算准确度的情况下,以提高计算效率为目的,设计新的语义关联度的计算方法。

我们提出了一种语义关联度的统一简化定义和计算方法。无论 v'_i 和 v''_i 是离散值集合还是连续值区间,都可以定义它们的交集和并集。定义 $r(v'_i,v''_i) = \dfrac{\| v'_i \cap v''_i \|}{\| v'_i \cup v''_i \|}$,其中, $\| \cdot \|$ 为某种范式,如当 v'_i 和 v''_i 为离散值集合时, $\| \cdot \|$ 表示的 COUNT、SUM、MAX、MIN、AVG 函数;当 v'_i,v''_i 为连续值区间时, $\| \cdot \|$ 可以表示欧氏距离。

显然,$0 \leqslant r(v'_i,v''_i) \leqslant 1$,当 $\| v'_i \cap v''_i \| = 0$ 时,$r(v'_i,v''_i) = r(v''_i,v'_i) = 0$;当 $\| v'_i \cap v''_i \| = 0$ 时,$r(v'_i,v''_i) = r(v''_i,v'_i) = 1$。

3 结语

本文提出了 OLAM 任务语义模型,利用这个模型可以分析 OLAM 任务的共性,为 OLAM 任务的表示、管

理和优化提供基础。在基于限制的语义片断的定义的基础上分析了任务语义模型中的基本操作。给出了语义片断语义关联度的一般计算公式。在大数据量的情况下,分析这些语义片断间的关系,对于优化查询、合理的缓存对象的确定均有重要的意义。

参考文献:

[1] 花文建,刘作良,杨 凡. 一种新的知识表示模型及其规则推导[J]. 空军工程大学学报(自然科学版),2003,4(6):44-47.

[2] Kanellakis P C, Kuper G M, Revesz P Z. Constraint Query Language[A]. Proc. Of 9th ACM Symp On Principles of Database System (PODS90)[C]. New York:ACM Press,1990. 315-332.

[3] Goldin D Q, Kanellakis P C. Constraint Query Algebras[J]. Constraints Journal, 1996, 1(1):45-83.

[4] 陈良刚,徐贵红. 区间约束数据库查询语言 ISQL[J]. 计算机研究与发展,2000,(1):28-31.

[5] 陈京民. 数据挖掘概念与技术[M]. 北京:电子工业出版社,2002.

(编辑:门向生)

An OLAM Semantic Task Model Based on the Constraints

DUAN Yan, CHENG Ying-lei

(The Telecommunication Engineering Institute, Air Force Engineering University, Xi'an, Shaanxi 710077, China)

Abstract: Through analyzing plenty of decision analytical algorithms, an OLAM semantic task model based on the constraints is presented. The model can be utilized for analyzing the common characteristics of OLAM task and extracting the elementary operations, the basis of which is the semantic slices. The definition of semantic slices based on the constraints and the relationship judging methods of the semantic slices are given. The relationship among the semantic slices is the foundation of optimizing the OLAM transactions.

Key words: OLAM; semantic task model; semantic slices