主题数据库规划合理性估计的数学公式*

王玉书 董丕明

(辽宁大学计算机科学技术系 沈阳 110036) (大连铁道学院计算机科学系 大连 116022)

摘要 本文简述了主题数据库的概念,给出了企业的抽象描述,实体的相关度、实体的紧密度的概念和企业主题数据库规划估计的数学计算公式.文中还给出了一些验证主题数据库规划估计的例子.

关键词 数据库,实体,主题数据库,系统模型,战略规划.

数据库的应用使计算机广泛地应用于各个方面,反过来应用本身对数据库的理论和应用技术又提出了更高的要求.只有对数据库进行合理地规划,才能使数据库应用系统发挥更高的效率、保持其长期性、稳定性和适应应用的多变性.Peter Chen 于70年代最先提出了关系数据库的模式设计的ER方法[1],后来在这个方法上又发展成信息模型化方法.[2]这些方法都是以实体或对象作为数据模型单位.对于一个大的企业,使用ER模型化可以得到几百,甚至几千个实体.直接使用这些实体作为企业数据规划的单位,分析员和企业的管理人员难于交流和理解企业的数据模型,无法规划企业的数据,要想做出企业的长远数据规划几乎是不可能的.对于这种情况,James Martin 提出数据战略性规划的概念.提出了2种类型数据库,应用数据库和主题数据库.[3]主题数据库是与企业的主题有关、与常规计算应用无关的那些长期稳定的数据组成的.例如,产品数据库是主题数据库,而发票、订单等不是.主题数据库(或称超级组)是由多个关于企业主题的实体组成的,与企业主题有关的每一个实体属于一个主题数据库.因此,主题数据库构成了与企业主题有关的所有实体上的一个划分.若干个主题数据库能够构成系统.主题数据库对于一个企业数据战略数据规划具有意义,即对于建立长期稳定的数据库以及数据的分布式管理都有意义.通常主题数据库包括的实体由功能分析的职能领域和处理的范围来确定.本文试图对主题数据库确定的合理性做出定量的估计,提出了基于活动—实体联系的企业的数据规划的合理性定量估计的一种数学方法.其一方面可以用来对几个主题数据库规划方案的合理性做出定量的比较,另一方面也能为主题数据库的自动规划算法提供依据.

---

* 作者王玉书,1944年生,副教授,主要研究领域为人工智能,管理信息系统.董丕明,1944年生,副教授,主要研究领域为计算数学.

本文通讯联系人:王玉书,沈阳 110036,辽宁大学计算机科学技术系

本文1996-09-02收到修改稿

1 企业模型

下面我们对企业的模型做一个简单的描述. 一个企业的活动可以由3个级别加以描述. 最高级是职能(Function),指企业活动的大的领域. 如对于一个中等规模的公司,业务规划、财务、产品规划、生产规划、生产、材料、人事等都是职能. 每一职能可以由若干个处理(Process)组成. 处理是比职能更小的活动方面. 例如对于产品规划职能,它由产品设计、产品定价、产品说明书管理等处理组成. 处理还可能细分,由若干个处理组成,即处理可以分为多个层次. 每一个最底层的处理由若干个动作所组成. 动作是完成一项简单事情并且认为没有必要细分的操作活动(Action),即完成一项任务不能再分的最小的活动. 我们也把动作称作活动. 例如,产品说明书管理中登记说明书作为活动,不要把登记编号、类别、内容等每一项看作活动. 因为登记编号单独来看是无意义的. 一个企业的构成除了上述的活动外,还有就是企业中的数据. 数据分为属性类,如员工的姓名、年龄等. 若干个联系密切的属性类构成的一个集合——看作一个整体——称作实体. 对于活动,其操作的输入和输出的数据集合称作数据组(Data Group).

我们用 Ent 表示一个企业. $F=\{f_1,f_2,\ldots,f_m\}$ 是企业中所有职能构成的集合. $P=\{p_1,p_2,\ldots,p_n\}$ 是企业中所有过程构成的集合. $A=\{a_1,a_2,\ldots,a_p\}$ 是企业中所有活动构成的集合. $E=\{e_1,e_2,\ldots,e_q\}$ 是企业中的所有与企业主题有关的实体构成的集合.

对于一个实体可能参与到一个或多个活动中,我们用 $S_a(e)=\{a_{e1},a_{e2},\ldots,a_{er}\}$ 表示实体 e 参与的活动的集合.

定义 1. 一个六元组 $Ent=\{F,P,A,E,S,T\}$ 是一个企业,其中 F,P,A 和 E 定义如上,$S=\{S_a(e_1),S_a(e_2),\ldots,S_a(e_q)\}$,$T$ 是 $F\cup P\cup A$ 元素构成的森林. F 中的每一个元素恰是一个树的根. P 中的一些元素是根树的分支节点. A 中的元素都是作为树的叶节点. 对于 P 中元素作成的节点,其子节点或者全部由 P 中元素组成,或者全部有 A 中元素组成. F,P 和 A 中元素在 T 中恰好出现1次.

该定义即表达了一个企业中的各种活动和活动之间的联系,也表达了活动与数据之间的联系.

定义 2. 实体的非空有限集合,我们称作超级组.

定义 3. 设 $Ent=\{F,P,A,E,S,T\}$ 是一个企业,$E_1,E_2,\ldots,E_k$ 均是 E 中元素组成超级组,如果 $E=E_1\oplus E_2\oplus\ldots\oplus E_k$($\oplus$ 表示直和),我们称 $G=\{E_1,E_2,\ldots,E_k\}$ 是企业 Ent 主题数据库的一个规划,每一个 E_i 都称作企业 Ent 的一个主题数据库.

目前虽然系统分析员和企业数据管理员能够依据企业活动的功能做出企业主题数据库的规划,但还有2个问题需要考虑. 一个问题是不同的系统分析组可能得到不同的企业主题数据库的一个规划,那么如何对它们做比较,比较的标准是什么,即如何对主题数据库规划进行质量评估. 另一个问题是企业主题数据库的规划除了完全由人确定外,能否依据算法自动地生成. 第2个问题的解决依据前一个问题的解决.

第1个问题的任务是设计主题数据库规划质量评估函数

$$R(Ent,G)$$

该函数的值域是[0,1],其中 $Ent=\{F,P,A,E,S,T\}$,$G=\{E_1,E_2,\ldots,E_k\}$,$E=E_1\oplus E_2\oplus\ldots\oplus E_k$. 当函数值越接近1时,表明规划得越好.

2 企业主题数据库规划的质量评估函数

对数据的实体与实体之间以及实体与活动之间的联系做定性和定量描述是用计算的方法确定主题数据库规划的前提. 实体和活动密不可分. 活动不能没有实体. 例如,货架上没有货,就没有卖货的活动. 反之,不参与任何活动的实体也是无意义的. 一种自然的想法是通过活动联系密切的实体应属于同一个主题数据库,这是我们的企业主题数据库规划的质量评估函数构思的出发点. James Martin 依据实体和活动之间的联系给出了2个实体相关度的定义.[3]我们以2个实体相关度定义为基础给出了其它一些定义.

设 S 是一个有限集合,我们用 $|S|$ 表示集合 S 的元素的个数,即秩数.

定义 4. 设 e 是一个企业中的实体,$S_a(e)$ 是 e 参与的所有活动构成的集合,我们把 $|S_a(e)|$ 称作实体 e 关于活动的相关数,简称实体 e 的相关数,记作 $N_a(e)$.

定义 5. 设 a 是企业中的一个活动,$S_e(a)$ 是 a 涉及的所有实体构成的集合,我们把 $|S_e(a)|$ 称作活动 a 关于实体的相关数,简称活动 a 的相关数,记作 $N_e(a)$.

设 $E=\{e_1,e_2,\ldots,e_q\}$,$A=\{a_1,a_2,\ldots,a_p\}$,$q\times p$ 阶矩阵 R 定义为

$$r_{ij}=\begin{cases}0, & \text{如果 } e_i \text{ 不参与活动 } a_j\\ 1, & \text{如果 } e_i \text{ 参与活动 } a_j\end{cases}$$

该矩阵称作实体活动矩阵,也称 E-A 表. E-A 表的一行元素的和是一个实体的相关数,一列元素的和是一个活动的相关数.

例1:表1是 E-A 表的例子.

表1　E-A 表

	a_1	a_2	a_3	a_4	a_5	a_6
e_1	0	0	1	1	0	0
e_2	1	1	1	0	0	0
e_3	0	0	0	0	1	0

定义 6. 设 e_1 和 e_2 是一个企业中的2个实体,我们把 $|S_a(e_1)\cap S_a(e_2)|$ 称作实体 e_1 和 e_2 的相关数,简记为 $N_a(e_1,e_2)$. 显然,$N_a(e_1,e_2)=N_a(e_2,e_1)$.

定义 7. 设 e_1 和 e_2 是一个企业中的2个实体,我们把 $N_a(e_1,e_2)/N_a(e_1)$ 称作实体 e_2 关于实体 e_1 的相关数,简记为 $a(e_1,e_2)$. 类似地,$a(e_2,e_1)=N_a(e_1,e_2)/N_a(e_2)$ 是 e_1 关于实体 e_2 的相关数.

一般说来,$a(e_1,e_2)\neq a(e_2,e_1)$.

定义 8. 设 $E_i=\{e_1,e_2,\ldots,e_k\}$ 是不同实体构成的集合. A_i 为 E_i 参与的活动的集合,即 $A_i=S_a(e_1)\cup S_a(e_2)\cup\ldots\cup S_a(e_k)$. 不妨设 $A_i=\{a_1,a_2,\ldots,a_t\}$,记 $E'_i=\{S_e(a_1)\cup S_e(a_2)\cup\ldots\cup S_e(a_t)\}$,即 $a_1,a_2,\ldots,a_t$ 涉及的所有实体构成的集合,把

$$\lambda_a(E_i)=|E_i|/|E'_i|$$

称作实体集 E_i 的活动相对数.

该数反映了一组实体与其有关的活动所涉及的实体比例.

定义 9. 设 $E_i=\{e_1,e_2,\ldots,e_k\}$是不同实体构成的集合,$e$ 是一个实体. 令

$$a(E_i,e)=\max\{\sum_{j=1}^{k}(a(e_j,e)\times N_a(e_j))/\sum_{j=1}^{k}N_a(e_j),\sum_{j=1}^{k}(a(e,e_j)\times N_a(e))/\sum_{j=1}^{k}N_a(e)\}$$

即
$$a(E_i,e)=\max\{\sum_{j=1}^{k}(a(e_j,e)\times N_a(e_j))/\sum_{j=1}^{k}N_a(e_j),\sum_{j=1}^{k}(a(e,e_j)/k)\}.$$

我们把 $a(E_i,e)$称作 e 关于实体集 E_i 的相关数. 当 $e\in E_i$ 时,$a(E_i,e)$称作 e 关于实体集 E_i 的内相关数;当 $e\notin E_i$ 时,$a(E_i,e)$称作 e 关于实体集 E_i 的外相关数.

实体与实体集的相关数反映了一个实体与一个实体集联系的密切程度. 该值越高,联系越密切,它用于反映一个实体与一个主题数据库的联系程度.

定义 10. 设 $E_i=\{e_1,e_2,\ldots,e_k\}$是不同实体构成的集合. 令

$$T_{upp}(E_i)=\max\{a(E_i,e_j)\mid j=1,2,\ldots,k\},$$

$$T_{low}(E_i)=\min\{a(E_i,e_j)\mid j=1,2,\ldots,k\},$$

$$T_{ave}(E_i)=\sum_{j=1}^{k}a(E_i,e_j)/k$$

它们分别称作实体集的最大、最小和平均紧密度. 平均紧密度也称作该实体集的紧密度.

实体集紧密度的概念反映了一个实体集中的实体通过它们的活动发生的联系的程度. 当实体集为一个主题数据库时,该实体集的最大、最小和平均紧密度分别称作该主题数据库的最大、最小和平均紧密度. 平均紧密度称作该主题数据库的紧密度. 特别地,当 E_j 中只有一个元素时,实体集的紧密度为 1. 这与我们的直观想象是一致的. 注意,每一个主体数据库紧密度高的规划从整体上看不一定是好的规划,一个好的规划必须从局部和总体的双方面来考虑.

定义 11. 设 $G=\{E_1,E_2,\ldots,E_m\}$是企业 Ent 的主题数据库的一个规划,我们令

$$R(E_1,E_2,\ldots,E_m)=\sum_{j=1}^{m}\lambda_a(E_j)\times|E_j|\times T_{ave}(E_j)/|E|$$

或简记为
$$R(G)=\sum_{j=1}^{m}\lambda_a(E_j)\times|E_j|\times T_{ave}(E_j)/|E|$$

其中 $|E|=|E_1|+|E_2|+\ldots+|E_m|$. 我们把函数 $R(G)$称作规划 G 的评估函数,R 的值域是$[0,1]$. 函数值 $R(G)$称作规划 G 的合理度.

R 值越大,规划越合理. 该评估函数的定义考虑规划中的 3 个因素:企业中实体与活动的联系每个主题数据库中实体参与的活动与企业中的有关活动;每个主题数据库中的实体在企业中所有实体中占据的比例.

3 规划评估函数的例子

下面给出检验规划评估函数使用的几个简单例子.

例 2:企业 Ent 的 E-A 表如表 2 所示.

表 2

	a_1	a_2	a_3	a_4	a_5
e_1	1	0	1	0	0
e_2	0	1	0	0	0
e_3	0	0	0	1	1

我们把 e_1,e_2 和 e_3 归并成一个主题数据库 $E=\{e_1,e_2,e_3\}$. 令 $G=\{E=\{e_1,e_2,e_3\}\}$. 计算得 $a(E,e_1)=0.4, a(E,e_2)=0.2, a(E,e_3)=0.4; T_{ave}(E)=0.33, \lambda_a(E)=1$, 于是 $R(G)=0.33$.

如果 $G=\{E_1=\{e_1\},E_2=\{e_2\},E_3=\{e_3\}\}$, 那么

$$T_{ave}(E_1)=T_{ave}(E_2)=T_{ave}(E_3)=1,\ \lambda_a(e_1)=\lambda_a(e_2)=\lambda_a(e_3)=1$$

于是 $R(G)=1$.

上面计算出的值表明, 通过活动不能联系起来的实体合并到一个主题数据库里的规划是不令人满意的.

例 3: 企业 *Ent* 的 *E-A* 表如表 3 所示.

令 $G_1=\{E_1=\{e_1\},E_2=\{e_2,e_3\},E_3=\{e_4,e_5,e_6\}\}$, 计算可得 $R(G_1)=1.0$.

令 $G_2=\{E_1=\{e_1,e_2\},E_2=\{e_3,e_4\},E_3=\{e_5,e_6\}\}$, 计算可得 $R(G_2)=0.4$.

从中可以看出 G_1 比 G_2 好.

例 4: 企业 *Ent* 的 *E-A* 表如表 4 所示.

令 $G_1=\{E_1=\{e_1\},E_2=\{e_2,e_3,e_4\},E_3=\{e_5,e_6\}\}$, 计算可得 $R(G_1)=0.5$.

令 $G_2=\{E_1=\{e_1\},E_2=\{e_2,e_3\},E_3=\{e_4,e_5,e_6\}\}$, 计算可得 $R(G_2)=0.5$.

令 $G_3=\{E_1=\{e_1\},E_2=\{e_2,e_3,e_4,e_5,e_6\}\}$, 计算可得 $R(G_3)=0.73$.

从中可以看出 G_3 的评估值比 G_1 和 G_2 都高. 这是由于 e_2,e_3,e_4,e_5,e_6 通过活动的联系比较密切(每 2 个实体之间至少有 50% 以上的活动是共同的), 把它们合在一起较为合理.

表 3

	a_1	a_2	a_3	a_4	a_5	a_6
e_1	1	0	0	0	0	0
e_2	0	1	1	0	0	0
e_3	0	1	1	0	0	0
e_4	0	0	0	1	1	1
e_5	0	0	0	1	1	1
e_6	0	0	0	1	1	1

表 4

	a_1	a_2	a_3	a_4	a_5	a_6
e_1	1	1	1	0	0	0
e_2	0	0	1	1	1	1
e_3	0	0	1	1	1	0
e_4	0	0	1	1	1	1
e_5	0	0	0	1	1	1
e_6	0	0	0	1	1	1

4　进一步的讨论

依据上面给出的概念和计算公式, 即实体和活动的关系, 可以设计一些对实体聚类算法, 即采用计算的方法产生企业的主题数据库的规划. 对于得到的不同规划使用评估函数做出数值评估. 企业的信息管理人员可以将计算得到的规划评估值较高的几个作为企业主题数据库规划的初稿, 再加以人工修改产生企业主题数据库的最终规划. 这种采用计算得到企业主题数据库规划的方法产生规划需要的时间少. 尤其对于实体和活动数量大、实体和活动

关系复杂以及信息规划人员经验少的情况,更有用武之地.我们认为这种建立主题数据库规划的方法值得进行深入研究.

参考文献

1 Martin James. Strategic data planning methologies. New Jersey: Prentice-Hall INC, 1982. 48～63, 134～138.

2 萨师煊等.数据库系统概论.北京:高等教育出版社,1984.19～23.

3 Peter Coad 等.面向对象的分析.北京:北京大学出版社,1991.13～23.

MATHEMATICAL FORMULA OF EVALUATING THE RATIONALITY OF SUBJECT DATABASES

WANG Yushu

(*Department of Computer Science and Technology Liaoning University Shenyang* 110036)

DONG Piming

(*Department of Computer Science Dalian Railways Institute Dalian* 116022)

Abstract In this paper, the authors state the concept about subject database, then give an abstract, description of enterprise model and present entity affinity, entity set affinity and the formula evaluating the rationality of enterprise subject databases. Some examples to verify the formula are shown. Finally, the purposes of the formula are pointed out.

Key words Database, entity, subject database, system model, strategic planning.