

OLAP系统

会话视图

用户

联机分析处理(OLAP)

计算机科学 2000 Vol. 27 No. 12

数据库

OLAP 系统中用户的会话视图

The Session View of Users in OLAP Systems

赵玉源 梁阿磊 白英彩

(上海交通大学金桥网络工程中心 上海 200030)

59-61

TP311.13

Abstract In OLAP systems, it may be dull for users to find the interesting information after many multi-dimensional operations. In fact, each user usually takes care of some information. The different sessions of every user may show some similarity. On the basis of up conclusions, the concept of session view of users is given in this paper. With it, user can pay his attention to some information that he is interested in and find the interesting information quickly.

Keywords OLAP, Session, View

1. 前言

在数据库系统中,视图是由基表或其它视图导出的虚表,只在数据目录中保留其逻辑定义,而不作为一个表实际存储在数据库中。数据库管理员可以为特定的用户或用户群定义一个或多个视图,以限制或引导他们对数据库的访问。用户也可以定义视图,把自己的视野集中在有意义的范围内。在文[1]中,提出了 Web 用户视图的概念,将 Web 用户的注意力限制在其感兴趣的范围内。

在数据仓库中保存了大量的信息,以对决策进行支持。此时,用户的查询主要是对大量数据的综合、聚集,一般都比较复杂。OLAP(联机分析处理)系统的出现,可以让用户以多维视图的方式,方便地对数据仓库进行访问。在 OLAP 系统中,用户可以从一个预先设定的入口点(通常是某一个查询的结果)开始,通过向上综合(roll-up),向下钻取(drill-down),旋转(pivoting)等多维操作[2],形成一系列的查询,找到自己感兴趣的东西。通常,用户与 OLAP 系统进行的由一系列查询构成的分析过程称为会话。

虽然,通过 OLAP 系统用户可以方便地访问数据仓库,但随着数据仓库中数据的增加,用户也不得不通过大量的多维操作,才能最终找到自己感兴趣的信息。此时,整个多维分析过程也变得枯燥、乏味了。而事实上,每个用户都关注其特定关心的内容。比如,企业的高层管理人员对关系整个企业的数据比较关心,而各部门的中、低层管理人员更关注于本部门的数据。这

样,某一用户与 OLAP 系统进行的不同会话通常就会有一定的相似性。基于此,本文引入了数据库中视图的概念,提出了 OLAP 系统中用户会话视图的概念。这样,就可以将用户的注意力集中在其所关心的范围内,使其能更好地找到自己感兴趣的数据。

为了方便对会话过程中用户的查询行为进行分析、比较,本文提出了一个多维数据模型;形式化地给出了 OLAP 系统用户会话视图的定义;并介绍了一个获取用户会话视图的方法和一个例子。

2. 多维数据模型

为了对 OLAP 系统中会话中用户的查询行为进行分析,必须对底层的多维数据进行形式化的描述。目前,已经提出了多个形式化的多维数据模型[3~6]。这里我们在文[6]的基础上,建立了一个能够有效支持用户查询行为分析的多维数据模型。

定义 2.1 一个维的层 l 是域 $dom(l)$ 中有限个元素的集合。 Ψ 表示多维数据所有层的集合。

定义 2.2 多维数据中的一个维 d 定义为一个二元组 $d = (H, \leq)$, 其中 H 是组成维 d 所有层的集合, $H \subseteq \Psi$, $\leq$ 是 H 上的线性序关系, 即 $\forall l_i, l_j \in H \rightarrow l_i \leq l_j, \forall l_i \leq l_j$ 。每一个维 d 都含有一个特殊的层 $d.all$, 它用来表示由所有基本数据聚集形成的层。

定义 2.3 一个多维数据模式是一个元组 $MD = (D, M)$, 其中 $D = \{d_1, \dots, d_k\}$ 是多维数据所有维的集合, $M = \{m_1, \dots, m_n\}$ 是所有指标的集合。

赵玉源 博士研究生, 主要研究领域为计算机网络、数据仓库技术。**梁阿磊** 博士研究生, 主要研究领域为计算机网络。**白英彩** 教授, 博士生导师, 主要研究领域为分布式系统、计算机网络。

函数 $bel(l): \Psi \rightarrow D$ 将一个层映射到所对应的维。

例 2.1 如图 1 所示的一个跨国汽车修理公司面向汽车修理主题的 ME/R 概念模型^[1], 用我们建立的多维模式可以表示如下:

$MD = (D, M)$

$D = \{d_{time}, d_{garage}, d_{vehicle}\}$

$d_{time} = \{ \{ day, month, year, d_{time_all} \}, day \leq month \leq year \leq d_{time_all} \}$
 $d_{garage} = \{ \{ garage, region, country, d_{garage_all} \}, garage \leq region \leq country \leq d_{garage_all} \}$
 $d_{vehicle} = \{ \{ vehicle, brand, d_{vehicle_all} \}, vehicle \leq brand \leq d_{vehicle_all} \}$
 $M = \{ quantity, costs(total) \}$

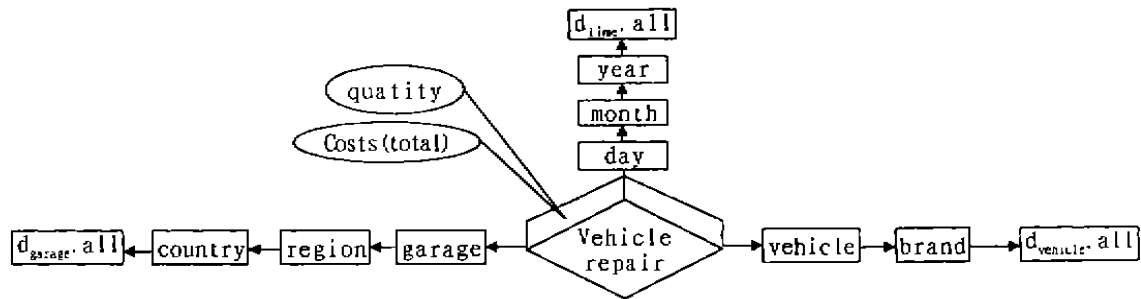


图 1 一个汽车修理公司的 ME/R 概念模型表示

3. OLAP 系统中用户的会话视图

在 OLAP 系统中, 设 D_q 是一个查询 q 涉及到的多维数据所有层组成的集合, $D_q \subseteq \Psi$, 则作为一个有效的查询, q 应该满足以下条件^[6]:

(1) $\bigcup_{l \in D_q} bel(l) = D$;

(2) $bel(l_1) \neq bel(l_2)$, 其中 $\forall l_1, l_2 \in D_q, l_1 \neq l_2$ 。

条件(1)、(2)确保了查询在每一个维中只涉及到一个层。本文所指的查询均为有效查询。

定义 3.1 在 OLAP 系统中, 如果两个查询 q_1 、 q_2 , 无论底层数据库的内容怎样改变, 其查询结果都完全相同, 那么我们就称 q_1 和 q_2 互为等效查询, 记为 $q_1 = q_2$ 。显然, 对于查询 q_1, q_2, q_3 , 如果 $q_1 = q_2, q_2 = q_3$, 则有 $q_1 = q_3$ 。

定义 3.2 为了找到一个商务问题的答案或为了验证先前的一个假设, 用户与 OLAP 系统进行的由一系列查询构成的分析过程称为会话。会话由构成会话的数个查询按时间先后形成的序列表示, 即 $S := \langle q_1, \dots, q_n \rangle$ (若其中包含有多个等效查询, 那么互为等效查询的多个查询用同一个符号表示), 其中, q_{i+1} 是由 q_i 的查询结果经过一个多维分析的基本操作(向上综合、向下钻取、旋转等)形成的 ($1 \leq i < n$)。

定义 3.3 会话 S 中含有查询的个数, 称为 S 的长度。

虽然在数据仓库中保存了大量的数据, 但通常用户都会有相对稳定的信息需求, 都有其特定关心的数据内容。这样, 就使得用户为找到其关心内容的会话之间, 有一定的相似性。此时, 如果有一个适当的判别规则, 就可以根据这个判别规则, 把用户与 OLAP 系统

所进行的所有会话过程按照其相似性的程度加以分类。这样当用户对某个相似类的兴趣强度超过了特定的阈值, 我们就可以把它定义为用户的一个会话视图。

定义 3.4 一个用户与 OLAP 系统的所有交互过程, 可以看作是用户执行的所有会话形成的序列(按时间先后), 我们称为会话序列, 可表示为: $ES := \langle S_1, \dots, S_n \rangle$ 。设 R 是一个给定的相似性判断规则, Γ 是 ES 关于 R 的相似类的集合, F 是一个从 Γ 到非负实数的函数(用来度量用户对相似类的兴趣强度), H 是一个给定的阈值, C 是 Γ 在关系 R 下的一个相似类, 若 $F(C) > H$, 则称 C 为该用户的一个会话视图。

4. OLAP 系统中用户会话视图的获取

4.1 相似性判别规则

相似性判别规则是用户会话序列上的一个关系(称为相似关系)。显然, 用户的每个会话与其本身是相似的。其次, 若某个会话 A 与会话 B 相似, 则会话 B 也应该与会话 A 相似。也就是说, 相似关系应该是自反的、对称的, 因而是一个相容关系。

由于同一会话中的所有查询都是用户为达到同一目标(找到一个商务问题的答案或为了验证先前的一个假设)形成的, 所以同一会话中的每个查询都应该有一定的相似之处。此时, 如果两个会话包含有相同或等效的查询, 那么我们有理由相信这两个会话也有一定程度的相似。此外, 如果会话 A 与会话 B 含有相同或等效查询, 会话 B 与会话 C 含有相同或等效查询, 那么即使 A 与 C 不含有相同或等效查询, A 与 C 也应该有一定程度的相似。这种相似性可以看作是由 A 与 B 和 B 与 C 的相似性传递得到的。所以, 我们有如下的

相似性判别规则:

设 A, B 是用户的两个会话, 我们说它们是相似的, 如果下面的条件(1)或(2)成立:

(1) 如果会话 A 与会话 B 含有相同或等效查询;

(2) 存在着会话 $A_0 = A, A_1, \dots, A_k = B$, 使得 A_{i-1} 与 $A_i (i=1, \dots, k)$ 含有相同或等效的查询。

4.2 兴趣强度度量

根据用户会话的相似性, 我们可以将用户所有的会话划分为不同的相似类。显然, 用户在同一个相似类的会话中, 有着相同或相近的信息需求。但是在不同的相似类之间, 用户的信息需求是不同的。同时, 用户对不同信息需求的强度也有所不同。那么如何度量用户的兴趣强度呢? 我们认为可以从两个方面来进行: (1) 如果用户的某个会话长度远远大于其它的会话长度, 我们可以认为用户在这个会话过程中花费了大量的精力; (2) 如果有许多会话都对应于某种信息需求, 那么可以认为这是用户的一个持续的兴趣所在。

相似类由彼此相似的会话组成, 与用户的某种特定的兴趣和信息需求相对应。所以, 我们可以综合考虑相似类中所包含的会话的个数和这些会话的长度来度量用户的兴趣强度。下面, 给出一个简单的计算公式:

$$F(C) = (l/L + n/N)/2 \quad (1)$$

其中 F 是用户兴趣强度的度量函数, C 是一个相似类, l 是相似类 C 的长度 (C 中包含的会话的长度之和), L 是用户所有相似类的长度的和, n 是 C 中包含的会话的个数, N 是用户所进行的所有会话的个数。从公式(1)可以看出, 一个相似类包含的会话越长、越多, 则意味着用户对它的兴趣越大、越持久。显然, 有

$$\sum_{C \in r} F(C) = 1 \quad (2)$$

4.3 阈值

由于会话视图是用户特别感兴趣的地方, 因此阈值一般应明显高于平均兴趣强度值。

4.4 视图的获取

给定相似性判别规则、用户兴趣度量函数和阈值, 就可以从用户的所有交互过程中得到视图了。下面我们给出一个简单的例子, 来说明用户会话视图的产生过程。

例 4.1 设 $ES = (S_1, S_2, S_3, S_4, S_5, S_6, S_7)$, 是某

用户与 OLAP 系统之间的所有会话形成的会话序列。其中, $S_1, S_2, S_3, S_4, S_5, S_6, S_7$ 分别由下面的查询序列构成: $(A, B, E, F, I), (B, A, B, D, G), (A, X, Y, Z), (X, W, T), (U), (M, O, N), (M, J, L)$, 则按照相似性判别规则, 所有的会话可以分成 3 个相似类: C_1, C_2, C_3 。其中,

C_1 包含 4 个会话过程: S_1, S_2, S_3, S_4 , 长度为 17;

C_2 包含 1 个会话过程: S_5 , 长度为 1;

C_3 包含 2 个会话过程: S_6, S_7 , 长度为 6。

ES 的长度为 $17+1+6=24$, 则

$$F(C_1) = (17/24 + 4/7)/2 = 0.64$$

$$F(C_2) = (1/24 + 1/7)/2 = 0.09$$

$$F(C_3) = (6/24 + 2/7)/2 = 0.27$$

可以知道用户的平均兴趣强度为: 0.33, 如果设定阈值为 0.5, 则 C_1 被选出来作为视图提交给用户。

结束语 本文通过对用户与 OLAP 系统的交互过程进行分析, 提出了用户会话视图的概念, 并给出了从中获取用户会话视图的方法。通过用户会话视图, 可以将用户的视野限制在有意义的范围内, 并进一步可以进行缓存管理、浏览引导等方面系统优化工作。

参考文献

- 1 阳小华, 周龙骥. Web 用户的视图. 软件学报, 1999: 691~693
- 2 王珊, 等. 数据仓库与联机分析处理. 北京: 科学出版社, 1998. 85~90
- 3 Li C, Wang X S. A Data Model for Supporting On-Line Analytical Processing. In: Proc. Conf. on Information and Knowledge Management. 1996. 81~88
- 4 Agrawal R, Gupta A, Sarawagi S. Modeling Multidimensional Database. In: Proc. of the ICDE. 1997. 232~243
- 5 Vassiliadis P. Modeling Multidimensional Databases, Cubes and Cube Operations. In: Proc. of 10th SSDBM. 1998. 53~62
- 6 Sapia C. On Modeling and Predicting Query Behavior in OLAP Systems. In: Proc. of the International Workshop on Design and Management of Data Warehouse (DMDW'99). 1999. 2.1~2.10
- 7 Sapia C et al. Extending the E/R Model for the Multidimensional Paradigm. ER Workshops. 1998. 105~115