复杂层次多维数据仓库中判断数据请求有解的一种方法

徐 强 顾毓清
(中国科学院软件研究所 北京100080)

A Method to Judge Whether a Data Request is Satisfiable in a Complex-Level-Structure Multidimensional Data Warehouse

XU Qiang GU Yu-Qing
(Institute of Software, Chinese Academy of Sciences, Beijing 100080)

Abstract The dimensions of many data sets have a complex-level structure in current multidimensional data warehouse models. But the data warehouse can't implement all the data at every node of these dimension levels because of the large amount of data, so it's necessary to judge whether a data request is satisfiable in a complex-level-structure multidimensional data warehouse. In this paper, a method is presented to solve this problem. The characteristics of prime number are applied.

Keywords Data warehouse, Multidimensional data model

1 多维数据模型概述

多维数据模型因为能够有效地支持联机分析处理(online analysis processing, OLAP)而引起了人们越来越多的注意。最近几年，人们提出了几种多维数据模型。这些数据模型把数据集合视为多维空间中的点集，把数据集合的属性分为维和度量两类。维属性用来描述度量属性，是多维空间的维度。度量属性的值用来做分析处理，是多维空间中的点。最初的模型不能表示维层次结构，进一步能够表达简单的维层次结构(即只有一条路径的层次结构)，后来能够表示满足具有代数格特征的维层次结构，最后多维数据模型发展到能够表示复杂维层次结构。所谓复杂维层次结构是指一个维的层次之间不是简单的单路径关系，而是构成了一个有向的无环图。图1给出的一个多维数据模型时间维各层次间的复杂关系。

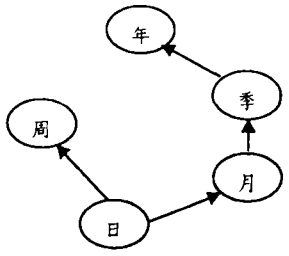

图1 多维数据模型中复杂层次关系

然而，由于数据量的问题，数据仓库并不能保证数据空间各个维度的每个层次组合结点上都有数据实现。高层次结点的数据可以从低层次结点的数据通过维集聚来获得。例如在一个商场数据仓库系统中某商场全年的销售情况可以从商场每月的销售情况汇总得到，但如果仓库中没有日销售记录，那么数据仓库就不能回答商场某天的销售情况。这就需要有一种方便的办法来判断一个数据请求在复杂层次的多维数据仓库中是否有解。

2 多维数据仓库模型

2.1 元数据

请求 Request 描述的是用户向数据仓库提交的数据访问请求，它由维度、粒度、度量等属性组成，可以理解为一个多维数据空间上的数据体。可以通过交、并、差、选择、衍生、维集聚等操作对请求进行加工、分解和演变。其中前5种操作不改变请求的维度空间，而维集聚操作可能会引起维度空间的变化。

维度 Dimension 刻画用户请求中数据空间中的各个维度，即数据体上的边，例如时间、地域等。每个维之间是相互独立的，不存在交叉。

粒度 Granularity 刻画用户请求中每个维度上的统计粒度。粒度必须依赖于某个维度才有意义，例如时间维度上的月、日等粒度，地域维度上的省、市等粒度。在同一维度的不同粒度间可能存在"偏序"关系，例如月可以由日汇总得到，年可以从月汇总得到等。这种"偏序"关系在维度的各个粒度之间构成了一个有向无环图。

度量 Measure 刻画用户请求中的数据，即数据体上内部的点，例如销售额、人口数、平均年龄等。度量必须描述在一定的维度上才有其意义，例如某商场的销售额，某省的人口数等。度量集合 Measure 通过函数 R 依赖维度(Dimension)：

$$R: DOM(d_1)\times DOM(d_2)\times\cdots\times DOM(d_n)\rightarrow DOM(m_1)\times DOM(m_2)\times\cdots\times DOM(m_k)$$

规则 Rule 与度量(Measure)相关的还有规则的概念，规则描述了度量之间的组合、衍生关系，比如人口数以及国民生产总值可以派生出人均国民生产总值度量。度量有直接度量(Direct Measure)和间接度量(Indirect Measure)两种，直接度量在物理元数据库中存在具体的数据项(Item)对应，而间接度量则需要通过 Rule 由直接度量衍生。一个度量可以既是直接度量又是间接度量，表示它既在物理元数据库中存在对应的数据项(Item)，也可以通过其他的度量衍生得到。

徐 强 硕士生，主要研究领域为数据仓库，软件工程。顾毓清 博士生导师，主要研究领域为软件工程技术。

$Request=\{r_1,r_2,\cdots,r_s\}$
$Dimension=\{d_1,d_2,\cdots,d_n\}$
$Measure=\{m_1,m_2,\cdots,m_k\}$
$Granularity=\{g_{11},g_{12},\cdots,g_{21},g_{22},\cdots,g_{n1},g_{n2},\cdots\}$
$(Dimension,Granularity)=\{(d_1,g_1),(d_2,g_2),\cdots(d_n,g_n)\}$
$Request=((Dimension,Granularity),Measure)$
$=(d_1,g_1,d_2,g_2,\cdots,d_n,g_n,m_1,m_2,\cdots,m_k)$

2.2 数据模型演算

请求的同构 Homotype　请求 $R=(d_{r1},g_{r1},d_{r2},g_{r2},\cdots,d_{rn},g_{rn},m_{r1},m_{r2},\cdots,m_{rs})$，$Q=(d_{q1},g_{q1},d_{q2},g_{q2},\cdots,d_{qn},g_{qn},m_{q1},m_{q2},\cdots,m_{qt})$同构，当且仅当两个请求的维度、粒度可以一一对应，即在同一个问题的空间。如果给数据仓库中的维度进行编号，按照维度编号的升序排列，则 R,Q 同构当且仅当 $d_{r1}=d_{q1},g_{r1}=g_{q1},\cdots,d_{rn}=d_{qn},g_{rn}=g_{qn}$。记为：$R\backsim Q$。请求的交、并、差等操作必须在同构的前提下才能进行，不同构的请求必须先通过维集聚操作转化成同构请求，才能进行交、并、差操作。

请求交 Intersect　如果请求 R,Q 同构，则：$R\cap Q=\{S|S$ 同时满足 R 以及 $Q\}$，$R\cap Q$ 仍然和 R,Q 同构，同时度量为 R,Q 度量的合并，也可以记为 INT(R,Q)。

请求并 Union　如果请求 R,Q 同构，则：$R\cup Q=\{S|S$ 满足 R 或者满足 $Q\}$，$R\cup Q$ 仍然和 R,Q 同构，同时度量为 R,Q 度量的合并，也可以记为 UNI(R,Q)。

请求差 Subtraction　如果请求 R,Q 同构，则 $R-Q=\{S|S$ 满足 R 同时 S 不满足 $Q\}R-Q$ 仍然和 R,Q 同构，同时度量为 R,Q 度量的合并，也可以记为 SUB(R,Q)。

请求的选择 Select　SEL(R,Precondition)＝{S|S 满足 R 同时其维度粒度度量满足条件 Precondition}，SEL(R,Precondition)的维度粒度度量均与 R 相同。

请求的衍生 Derive　DEV(R,Rule)＝{S|S 维度粒度均与 R 相同，同时度量通过 $Rule(d_1,g_1,\cdots,m_k)$演变得到}。

请求的维集聚 Collection　$R=(d_1,g_1),(d_2,g_2),\cdots,(d_n,g_n),m_1,m_2,\cdots,m_k$，假设粒度 d_i 上存在偏序 $g_i\rightarrow q_i$，即粒度 g_i 可以归并到粒度 q_i，例如：时间维上日粒度可以归并到月粒度，则：$COL(R,d_i,g_i,q_i)=\{S|S$ 的其他维度粒度均与 R 相同，(d_i,g_i)归并为(d_i,q_i)，相应的度量也进行 sum，avg…等归并操作。逻辑上 $COL(R,d_i,g_i,q_i)$是在统计粒度上的合并，引起了统计粒度的变粗，也有可能引起某个维度的简化，从而改变了请求的维度空间。

2.3 辅助命题

层次偏序关系　多维数据仓库中，有如下维度 $D=\{d_1,d_2,\cdots,d_n\}$，对于其中的维度 d_i，有如下的层次：$G_i=\{g_{i1},g_{i2},\cdots,g_{ik}\}$。在层次集合中，各层次之间有“直接偏序”关系$\leqslant$。$g_{iu}\leqslant g_{it}$表示层次 g_{iu}可以直接归并到层次 g_{it}。如果 $g_{is}\leqslant g_{it}\leqslant\cdots\leqslant g_{iu}$，则 g_{is}和 g_{iu}之间属于“偏序”关系，记做 $g_{is}<g_{iu}$。“直接偏序”关系使得层次之间构成了一个有向无环图。则数据请求可以格式化为：$R=(d_1,g_{i1},d_2,g_{i2},\cdots,d_n,g_{in})$。

命题　一个数据请求 $R=(d_1,g_{11},d_2,g_{22},\cdots,d_n,g_{nn})$有解，等价于：多维数据仓库中有直接数据源 R，或者存在问题 $R'=(d_1,g_{11},d_2,g_{22},\cdots,d_i,g_{iu'}\cdots,d_n,g_{nn})$，R′有解并且有 $g_{iu'}<g_{iu}$。

证明：充分性，若存在直接数据源 R，则显然数据请求 R 有解；如果存在问题 $R'=(d_1,g_{11},d_2,g_{22},\cdots,d_i,g_{iu}\cdots,d_n,g_{nn})$，R′有解并且有 $g_{iu'}<g_{iu}$，则通过维集聚操作，$R=COL(R',d_i,g_{iu'},g_{iu})$，请求 R 有解。必要性，反证如果不存在直接数据源，也不存在 R′，则意味着 R 无法通过维集聚的方式得到解，由于其他5种演算均不能改变维结构，所以 R 亦无解，故必要性成立。命题提供了判断数据请求 R 是否有解的基本解决方案，就是判断有没有可以归并到 R 并且有解的 R′存在。

3 判断算法实现

在此多维数据模型的基础上，利用素数的性质，来刻画层次间的偏序关系。假设多维数据仓库中，有如下维度 $D=\{d_1,d_2,\cdots,d_n\}$，对于其中的维度 d_i，有如下的层次：$G_i=\{g_{i1},g_{i2},\cdots,g_{ik}\}$。则为每个层次 g_{ij}分配一个数字 f_{ij}用于刻画层次 g_{ij}的特征，f_{ij}称为该层次的特征数，特征数构成的集合 F 称为特征数集合。

通过下面的标记算法 Marking algorithm 初始化模型：

```
Marking algorithm(Gi)
    for j=1 to k {
        pij=new_prime();
        fij=pij;
    }
    while Gi<> NULL{
        g0=find_root(Gi);
        for all gij∈Gi and gij<> g0 if g0≤gij fij=lcm(fij,f0);
        delete(Gi,g0);
    }
```

函数描述：

new_prime()：获得一个与以前的素数不同的新的素数，实现略。
find_root(G_i)：在 G_i 中找到一个根结点 g_0，即不存在 g_i 满足 $g_i\leqslant g_0$，实现略。
lcm(f_{ij},f_0)：取最小公倍数。
delete(G_i,g_0)：把 g_0从 G_i 集合中删除。
结论：$g_{ia}\leqslant g_{i1}\leqslant g_{i2}\leqslant\cdots g_{ib}(g_{ia}<g_{ib})$等价于 $f_{ia}|f_{ib}$。

当向数据仓库某维度添加新层次时，使用添加算法 Adding algorithm(G_i,g_0)：

```
Adding algorithm(Gi,g0)
    f0=p0=new_prime();
    for j=1 to k if gij≤ g0 f0=lcm(f0,fij);
    for j=1 to k if g0≤ gij{
        or m=1 to k and m <> j if fij|fim fim=lcm(fim,f0);
        fij=lcm(fij,f0);
    }
    add(Gi,g0);
```

函数描述：

new_prime()：获得一个与以前的素数不同的新的素数，实现略。
add(G_i,g_0)：向层次集合 G_i 添加新层次 g_0。
lcm(f_0,f_{ij})：取最小公倍数。

当从数据仓库删除某维度的某层次时，使用删除算法 Deleting algorithm(G_i,g_0)：

```
Deleting algorithm(Gi,g0)
    for j=1 to k if p0|fij fij=fij/p0;
    delete(Gi,g0);
```

函数描述：

delete(G_i,g_0)：把 g_0从 G_i 集合中删除。

层次 G 集合的特征数集合 F 非常直接、方便的刻画了层次间的归并关系。即：$g_i<g_j=f_i|f_j$，这为判断一个数据请求是否可以从另外一个请求归并得到提供了有效依据。

4 算法分析

以上面的时间维层次为例，假设：g_1为日，g_2为月，g_3为年，g_4为周，$p_1=2,p_2=3,p_3=5,p_4=7$。假设还有一个季层次 g_5，$p_5=11$。

算法有效性：

Marking algorithm(G)的执行结果：$f_1=2,f_2=6,f_3=30,f_4=14$。

Adding algorithm(G,g_5)的执行结果：$f_1=2,f_2=6,f_3=330,f_4=14,f_5=66$。

Deleting algorithm(G,g_2)的执行结果：$f_1=2,f_3=110,f_4=14,f_5=22$。

特征数集合 F 始终维持 $g_i<g_i=f_i|f_i$ 成立。

时间复杂度：在特征数集合 F 的支持下，每次判断数据请求之间是否有归并关系的时间复杂度为 O(1)；反之，需要搜索这个维度的所有层次，时间复杂度为 O(n)。

进一步工作 本文主要介绍了如何判断复杂层次多维数据仓库中数据请求是否有解，目的是为了建立一个可实现的基于查询优化的虚拟数据仓库模型，在实际的虚拟数据仓库建模中，还存在诸如用户请求解决方案的选择和自动生成，虚拟元数据与物理元数据的映射匹配等问题。目前我们正在本文所提出的方法的基础上实现面向 Web 的数据查询优化和处理方法，并进一步实现一个基于查询优化的虚拟数据仓库模型。

参 考 文 献

1 李建中，高宏. 一种数据仓库的多维数据模型. 软件学报，2000，11(7)：908～917

2 Bischoff J. Data Warehouse - Practical Advice From The Experts. 1998. 196～202

3 A model-driven architecture for Distributed Information Integration. http://www.omg.org

4 Inmon W H. Building the Data Warehouse Second Edition. 2000. 116～143

（上接第8页）

因此说，移动主机的移动大多数情况下是在同一层级的基站下移动的。基于上述两点，论文提出了一种限级微观移动协议。

5.2 协议基于的框架

这个协议基于 Cellular IP 框架结构，之所以选择 Cellular IP，因为 Cellular IP 框架中的实体在执行协议是用特殊的功能模块，Cellular IP 有更好的扩展性，并且它更好地考虑了实际的网络框架。

此外，最重要的一点是它天然地支持 Paging 机制，而论文中提出的协议就是利用 Paging 机制来实现的。

5.3 协议具体描述

要保证报文不丢失有两种方法：一种是在固定节点保留大量的报文，作为上述情况(3)(4)的缓冲；另外一种就是预先检测，并预先建立某种连接，使报文不致丢失。第一种情况不太实际，因为固定节点不知道缓存多少报文，并且这会占用固定节点大量的资源。因此我们选择后一种方法。

首先，移动主机检测到有可能发生基站的切换，这种检测要比目前的检测基站要宽，如检测移动主机接收信息的强度低于某一阈值比目前的大，在 overlap 期间接收到新基站广播后立即触发协议执行等等。换句话就是最大可能地预测移动主机的切换。

其次，在检测到最大的切换可能性之后，移动主机发送最低级 Paging 机制触发报文，该报文的作用是让最低级的路由器（基站的上一层路由器），建立对某一特定主机的 Paging 机制，这个机制的建立可以在路由器中修改路由 Cache 中转发表的状态。在转发表状态改变后，到移动主机的报文都会陪 Paging 到所有的下层基站。

最后，转发表状态的恢复需要移动主机确定其真正的完成切换之后，发送特殊类型的报文到路由器上。

5.4 协议中涉及到的技术和问题

(1)切换预测技术；(2)移动主机确认完成切换问题；(3)协议增加了报文的浪费，因为路由器要复制报文；(4)协议并没有太多增加路由器的处理负载，和增加路由器空间资源，因为转发表在 Cellular IP 中就已经有了。

小结 本文详细分析了仿真数据，并就数据显示的内容比较了我们所讨论的三种微观移动协议。在分析数据的基础上，本文还提出了一种改进的微观移动协议。由于 Internet 的无线访问这一趋势决定了微观移动协议这一领域的研究将越来越成为网络研究的一个热点，所以论文的研究工作会对以后的研究起着积极的作用。

参 考 文 献

1 Perkins C, et al. RFC 2002, IP Mobility Support, Oct. 1996

2 Gustafsson E et al. Mobile IP Regional Registration . draft - ietf - mobileip - reg - tunnel - 03. txt

3 Campbell A, et al. Cellular IP, draft-ietf-mobileip-cellularip-00. txt, Dec. 1999

4 Campbell A, et al. Design, Implementation, and Evaluation of Cellular IP. IEEE Personal Communications, Aug. 2000

5 Ramjee R. HAWAII: A Domain-based Approach for Supporting Mobility in Wide-area Wireless Networks

6 Ramjee R. Porta T L. IP micro-mobility support using HAWAII. draft - ietf - mobileip - Hawaii - 00. txt, 25 Jun 1999

7 Das S. TeleMIP: Telecommunications-Enhanced Mobile IP Architecture for Fast Intradomain Mobility. IEEE Personal Communications, Aug. 2000

8 Eardley P, et al. A Framework for the Evaluation of IP Mobility Protocols

9 Campell A. IP Micro-Mobility Protocols. COMET Group, Center for Telecommunications Research Columbia University

10 Ramjee R. IP-Based Access Network Infrastructure for Next-Generation Wireless Data Networks. IEEE Personal Communications, Aug. 2000

11 Mihailovic A, Shabeer M. Multicast For Mobility Protocol (MMP)For Emerging Internet Networks

12 Mihailovic A, Shabeer M. Sparse mode multicast as a mobility solution for internet campus networks

13 Welch B. Practical Programming in Tcl and Tk, Created: May 9, 1994 -bookTOC. doc-Copyright Prentice Hall-DRAFT: 1/13/95

14 The ns Manual,

15 Nam Manual

16 Xgraph Manual

17 NS Source Code. http://www.comet.columbia.edu/micromobility

18 NS tutor,

19 许小刚. NS 系统总结文档

20 Mohan S, Jain R. Two User Location Straegies for Personal Communications Services. IEEE Personal Communications, 1994

21 Chen Yi-an, A Survey Paper on Mobile IP

22 赵阿群，等. 网络层支持主机移动的研究. 计算机科学

23 赵阿群. 无线 Internet 研究进展

24 许小刚. 微观移动协议研究：[研究生毕业论文]. 2002-3-3