用关联维构建和操作在群体决策中的应用

李 晖 张世栋
(山东大学计算机科学与技术学院 济南 250100)

摘 要 为了解决传统数据仓库无法支持群体决策的问题,提出了一个支持群体决策的数据仓库中建立关联维的算法以及对于关联维的实际操作。通过在电力营销数据仓库系统中的实际应用,证明了算法的正确性和操作的可行性。表明,该方法可以得到更加高效的查询,提高与群体决策相关的辅助企业决策水平。

关键词 关联维,群体决策,数据仓库,数据操作

Construct and Operation of Associated Dimension in Group Decision

LI Hui ZHANG Shi-Dong
(Sch. Computer Sci., Shandong Univ, Jinan 250100)

Abstract To realize data warehouse supporting group decision, an algorithm used to build associated dimension is provided, meanwhile a set of corresponding data manipulation commands is studied. Experiments in electric power marketing data warehouse system indicate that this algorithm has retionalization, correctness, feasibility and high query efficiency. It also proves that this method has improved decision level of enterprises related to group decision.

Keywords Associated dimension, Group decision, Data warchouse, Data manipulation

目前,企业决策通常需要由群体做出,群体通常可以做出更有效的决策。在群体决策过程中,决策者们可以相互影响,拥有更多的决策模式,因此群体做出的决策通常更加正确和全面。

群体决策需要全方位的、充足的信息来支持决策的制定,这些信息中有些可能来自于分布的和异构的数据。传统的联机分析处理命令和操作[1]都是基于单个数据仓库的,它们不支持在多个数据仓库中操纵数据,所以很难支持群体决策。为此,我们建立了一个支持群体决策的数据仓库系统,并建立了相应的数据操作。它不仅兼容单个数据仓库操作命令,还能更好地支持群体的辅助企业决策。

本文将首先阐述关联维的相关定义,然后给出建立关联维的规则和建立关联维的算法,再介绍如何使用基于关联维的数据操作替代传统的联机分析处理操作,并通过电力营销的实例加以说明,最后提出将来的研究方向。

1 关联维定义

维是数据仓库中分析数据的角度,传统数据仓库中有多个维。进一步分析,会发现其中某些维是冗余的。造成这一现象的原因是数据仓库提供的数据是集成数据。

关联维$=\{D|\Phi(D)\}$,其中 D 表示关联维,$\Phi(D)$是构建关联维公式。

$D=D_1 \cup D_2 \cup \cdots \cup D_n$

即一个关联维是由数据仓库中的多个维或者关联维组成的集合。采用关联维的优势是用户可以从多个角度、多个侧面查询数据仓库中的数据,从而深入了解包含在数据仓库中的信息。比如在电力营销数据仓库系统中,既可以按照用电地区和用电类别查询销售量,也可以按照用电日期和用电地区进行查询,从而为群体决策提供有力的依据。

本文的应用中只考虑关联维,关联维从事实中发现数据(如图 1 所示)。

2 构建关联维方法

为了提出建立关联维的算法,先利用一个分析售电量的实例加以说明。实例中有一个售电量事实表,三个维表分别为日期维 d、地区维 l、用电类别维 k。其中日期维又分为三个维:年、月、日,地区维分为两个维:城市、省。用 dk 表示日期和用电类别的关联维,用 kl 表示地区和用电类别的关联维,用 dl 表示日期和地区的关联维,用 dkl 表示日期、地区和用电类别的关联维。将所有的维和关联维形成一个偏序集,如图 2 所示。

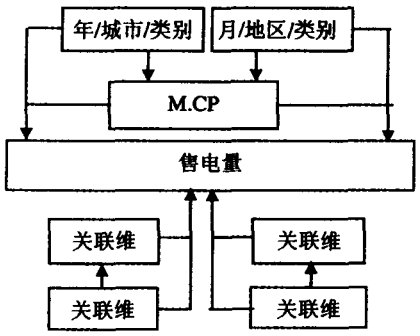

图 1 事实和关联维的关系

数据仓库中的查询是对集成数据进行的查询,它不涉及具体的、细节的数据。因此,多数情况下需要在更高的层次上建立关联维。因为高层次的关联维是建立在低层次的维的基础上的,可以认为在同样的维基础上建立较高层次关联维的概率和建立较低层次关联维的概率相等。例如,建立关联维

李 晖 博士研究生,主要研究领域为软件与数据工程。

2003年/山东的概率和建立维2003年8月/山东济南的概率相等。

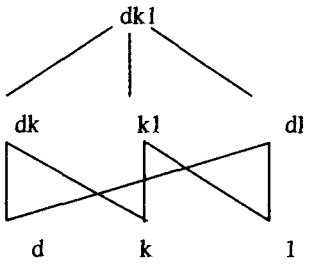

图2 维的偏序关系

本文的例子是针对电力营销相对具体的需求，通过统计用户的查询概率，得到如下的假设：假设：对用户需求q，查询出现的概率为 f_q。

基于这个假设，提出算法1用于建立最底层的关联维。

算法1：

输入：Q 是常用的复杂查询的集合、查询的概率 f_q 以及一个关联维的偏序集

输出：CV 是候选组合的集合和选中的候选组合的概率

setp 1：将 CV 设置为空集

step 2：对每个 q 在 Q 中作

(2.1) 对每个查询 q，找到所有属性 $a_1, a_2, \cdots, a_n$，每个属性都是一个维

(2.2) $CV=CV+\{a_1, a_2, \cdots, a_n\}$

(2.3) $f(a_1, a_2, \cdots, a_n)=f(a_1, a_2, \cdots, a_n)+f_q$

step 3：返回 CV, f

实际工作中用到的关联维查询及其概率见表1。

表1 常用关联维查询及概率

关联维查询	概率
2003年1月，济南	0.2
2002年，山东	0.1
2003年1月，济南	0.2
2003年1月1日，济南	0.0001
山东，居民用电	0.1
2002年，居民用电	0.2

表1中的关联维查询和概率作为算法1的输入，结果如下：

CV={(日期，地区)，(地区，用电类别)，(日期，用电类别)}

f(天，地区)=0.3001

f(天，用电类别)=0.2

f(地区，用电类别)=0.1

建立一个关联维时，必须考虑其维护、存储和查询的代价，因此需要根据用户查询的频率选择关联维。在实际应用中，因为不需要考虑存储的代价，所以维护关联维的代价与关联维的大小无关，只与关联维的数目有关。在关联维数目固定的情况下，认为建立 K 次关联维的代价和维护 K 次关联维的代价相同。得到如下结论：

只考虑算法得出的关联维的概率。

基于上述结论，结合下面的算法，得到最有效的关联维集合。

假设在系统中，允许建立的关联维的数目是 K，将集合 CV 中的关联维按照概率大小排序。在建立的相应关联维中，第一个关联维是最低层次的。建立关联维的规则如下：如果关联维对应的用户查询概率很小，并且查询概率相对于那些基于底层维的较高层次的关联维很小，这时就只能建立最低层的关联维，因为高层次的关联维可以根据底层的关联维创建。

算法1中，可以得到最低层的关联维。但是雪花模型中存在层次关系，根据上述规则，提出另一个算法用来选择合适的关联维。

算法2：

输入：算法1的结果 CV、用户常用查询集合 Q 和相应的查询概率 f_q

输出：需要建立的关联维集合 MV

setp 1：清空 MV，用来存储得到的关联维

对每一个关联维 v 和每一个在 v 上的高层次的关联维，设置 $f_v=0, i=0$，表示关联维的数目。

step 2：对每一个 CV 中的 v

{

对每一个在 Q 中的 q

{

找到查询 q 的所有属性 $a_1, a_2, \cdots, a_n$，其中每个属性都是维

如果 v 属于 QV，对 v_i 到每一个高层次的关联维在 v，$f_{vi}=f_{vi}+f_q$

}

}

step3：对每一个在 CV 中的 v

{

如果 $i>=K$，算法结束，返回 MV

F_z=关联维 v_i 总的概率，v_i 是包含 v 的高层次的维

如果 $(f_v-f_z)/f_v>m$

{

将 v 加入到 MV

$i++$

}

对每个 v 上的高层次维 v_i，

如果 $(f_v-f_{vi})/f_v>m$

如果 $i<K$

{

将 v_i 加入到 MV

$i++$

}

否则，算法结束，返回 MV

step 4：返回 MV

这样就得到了高层次关联维的集合。在算法中 m 是一个经验值。举例：输入算法1的结果得到算法2的结果，m 的值是0.001，结果是 MV={(年.月，城市)，(年.月，省)，(年.月，用电类别)，(省，用电类别)}

3 基于关联维的数据操作

关联维构建完成后，可以在信息查询中加以使用。以下的所有操作都是以电力营销数据仓库系统为例。每一种操作都是基于关联维的操作。

(1)切片

切片是在单个数据仓库中，从众多维中选择某些维为选择的维赋值，然后在数据方体中得到一片。在电力营销数据仓库系统中，具体操作如下：Eselect(时间：年=2003，月=1；工业用电)，查询2003年1月工业用电的销售量。

(2)切块

在单个数据仓库中选择某几个维，把一个维的值限制在一定范围中，也就是在一个数据方体中切下一块。在电力营销数据仓库系统中，这个操作和切片操作类似，操作如下：Eselect(时间：年=2003，月=between(1,6)；工业用电)，是在数据仓库中查询2003年1月到6月的工业用电销售量。

(3)上卷和下钻

在单个的数据仓库中，上卷意味着查看更高层的数据，下钻意味着看更具体的数据。这两个概念是相对的。上卷的目

的是查询更全面的数据，下钻是查询更详细的数据。具体操作如下：through(时间：年＝2003，月，NULL)，从按年组织的数据出发，察看 2003 年各个月的电力销售量。

(4)历史同期比较

对于多个数据仓库，历史同期比较操作提供了同以前数据进行比较的功能。操作如下：h_comparison(2001，2002，1，month，sum，NULL，GRAPH)，将 2001～2002 年每年的售电量按照月份进行比较，并以直方图的形式显示。

(5)横向比较

对于多个数据仓库，横向比较操作可以比较用电类别之间、区域之间的销售数据。操作如下：

h_comparison({工业用电，居民用电}，sum，时间：月，地区：城市)，将每个城市每月居民用电和工业用电的销售量进行比较。

从上面的实例中看出，电力营销数据仓库系统中的操作能实现传统的联机分析处理操作所实现的功能，所以可以使用上述数据操作命令替代传统的联机分析处理操作。

结论 本文以电力营销的实际需求为背景，根据群体决策的特殊要求，提出了一种用于构建关联维的算法以及在关联维上的操作。本算法能够帮助用户准确地把握市场趋势，作出更加合理的决策。

将来的研究工作和发展方向是基于混合数据仓库对象模型和雪花模型，以及如何加速查询的响应时间。

参 考 文 献

1 郑永清，高爱强，李庆忠，等．SCDDWS 中的数据操纵命令集[A]．12 届 CAD/CG 学术会议论文集[C]．贵阳：中国计算机学会，2002，497～501

2 隋琪，郑永清，杨少军，等．一个对象和雪花模型混合的数据仓库模型[A]．12 届 CAD/CG 学术会议论文集[C]．贵阳：中国计算机学会，2002，562～566

3 孙晓，李庆忠．支持企业群体决策的一种数据仓库模型[J]．计算机集成制造系统，2003，9：85～90

4 Codd E F，Codd S B，Salley C T. ProvidingOLAP(On-Line Analytical Processing) to User Analyst：An IT Mandate. [EB/OL]http://www.arborsoft.com/OLAP.html，2004

5 Chauduri S，Dayal U. An overview of data warehouseing and OLP technology[J]. SIGMOD Record，1997，26(1)：65～74

6 Gray J，Chaudhuri S，Bosworth A，et al. Data cube：a Relational aggregation operator generalizing group-by，cross-tab and sub Totals[J]. Data Mining and Knowledge Discovery Journal，1997，1(1)：29～53

(上接第 135 页)

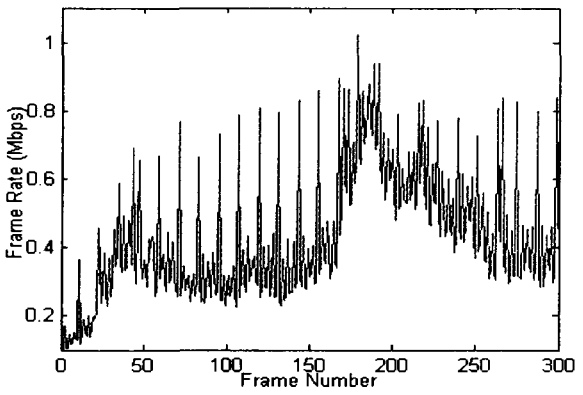

图 3 实验数据帧速率统计图

4.2 参数设置

1)对于(9)中的各影响因素的权，根据各因素的重要性，我们做如下设置：$\alpha(t_k)\equiv 10$，$\beta(t_k)\equiv 5$，$\gamma(t_k)\equiv 0.05$，$\lambda(t_k)\equiv 0.05$。

2)关于图 1 所示系统各参数设置如下：$T=5000$Btyes，$\hat{T}=10000$Bytes，$Q=6000$Bytes，$\tau=0.04$s，$C=0.3$Mbps，0.4257Mbps，1.0Mbps。

3)神经网络为 2 层的前馈神经网络；输入个数为 4；第一层神经元个数为 3，激活函数为 tansig；第二层为神经元个数 1，激活函数为 poslin。

4)粒子群算法中迭代次数为 200，惯性权重由初始值 0.9 到 0.1 线性递减。

4.3 实验结果及其分析

对不同的 C 的取值，分别在 $u=1572$（小于平均速率），$u=2231$（平均速率）、$u=5358$（峰值速率）和 u 由神经网络及粒子群算法控制下做了不同的实验，得出的数据如表 2。

表 2 实验所得 $J(u)$统计

u \ C	0.3 Mbps	0.4257 Mbps	1.0 Mbps
1572	1165200	1119100	1119100
2231	553580	443110	422300
5358	463540	39604	18
NN 和 PSA 控制	139930	22795	18

从表 2 我们可以看出，在固定复用器输出速率 C 的情况下，由神经网络和粒子群算法控制优化控制时，系统的代价函数值明显小于其他情况，从而证实了我们所提出方案的有效性。

结论 本文在给出了一个 MPEG 视频传输模型的基础上，详细给出了其系统参数说明和状态方程，以及评价系统性能的代价函数，提出了通过神经网络和粒子群算法对令牌桶产生速率进行控制的方案，并且通过实验证实了该方案的有效性，且能够显著地提高系统的传输性能。

参 考 文 献

1 IETF. Specification of the Controlled-Load Network Element Service[S]. RFC2211，1997

2 IETF. Specification of Guaranteed Quality of Service [S]. RFC2212，1997

3 IETF. An Architecture for Differentiated Services[S]. RFC2475，1998

4 IETF. Two-bit Differentiated Services Architecture for the Internet[S]. RFC2638，1999

5 Alam M F，Atiquzzaman M，Karim M A. Effects of source traffic shaping on MPEG video transmission over next generation IP networks[C]. In：IEEE Eight Intl. Conf. on Computer Communications and Networks，1999. 514～519

6 Lombaedo A，Schembra G，Morabito G. Traffic specifications for the transmission of stored MPEG video on the Internet[J]. IEEE Transactions on Multimedia，2001，3：5～17

7 Kennedy J，Eberhart R. Particle swarm optimization[C]. IEEE Intl. Conf. on Neural Networks，1995，4：1942～1948

8 Eberhart R. A new optimizer using particle swarm theory[C]. In：Proc. of the Sixth Intl. Symposium on Micro Machine and Human Science，1995. 39～43

9 Shi Y，Eberhart R. A modified particle swarm optimizer[C]. IEEE Intl. Conf. on Evolutionary Computation，1998. 69～73

10 http://www.tkn.tu-berlin.de/research/trace/trace.html