文章编号:1001 - 2486(2006)04 - 0083 - 05

维度汇总性问题及其对策*

陆昌辉,刘青宝,邓 苏,张维明
(国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:在联机分析处理中,为了提高查询的响应速度,预聚合是一种常用的方法,但在已有的研究中,关于维度汇总性的研究还相当少。从维度汇总性的基本概念及其分类出发,对维度汇总性的判断方法进行了研究,最后还给出了处理维度汇总性问题时的一些对策。

关键词:数据仓库;OLAP;汇总;分类;判断方法;对策
中图分类号:TP391 文献标识码:A

The Question about Dimensional Summarizability and Its Countermeasures

LU Chang-hui, LIU Qing-bao, DENG Su, ZHANG Wei-ming
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: In OLAP application, aggregation is the usual technique to improve the response speed of the users' query. However, there have been few attempts to completely characterize the ability to summarize measures along some dimensions. Based on this situation, this paper firstly gives some basic definitions about the dimension summarizability, and represents its taxonomy. Then it describes the technique of judging the summarizability in detail, and finally, its countermeasures are introduced.

Key words: data warehouse; OLAP; summarizability; classify; judging method; countermeasure

在数据仓库和联机分析处理系统中,为了提高查询的响应速度,精确高效的汇总数据是必不可少的[7]。汇总性的概念是在研究统计对象与 OLAP 维中的聚合导航时提出来的[2]。正如最初所说的,汇总性是指一个简单的聚合查询(通常称作汇总或者合并)能否根据其他预先计算的立方体视图计算得出。但是在某些情况下,如果不经过判断,就对度量值沿着某些维进行汇总操作,可能会得出不准确的结果,从而引发错误的决策。因此,判断维度汇总性问题是至关重要的。文献[3]指出了汇总性"是 OLAP 领域中被普遍忽略的一个重要论题",本文基于这样一个研究背景,对维度汇总性问题进行了研究。

1 维度汇总性的定义

定义 1(立方体视图) 定义在维 d 中层次 l 上的立方体视图可以用 $cv(d,F,l,af(m))$ 来表示,其中 F 是包含 d 中基础层次 l_{base} 的事实表;af 为一聚合函数;m 是 F 的一个度量。立方体视图 $cv(d,F,l,af(m))$ 描述了下列聚合视图:$\Pi_{l,af(m)}(F \triangleright \triangleleft (\Gamma^{l}_{l_{base}}(d)))$(其中 Π 表示投影操作,它并没有消除结果集中的重复元组;$\Gamma^{l}_{l_1}$ 为维层次 l_1 与 l_2 间的上卷操作)。

定义 2(维度汇总性) 给定一个维 d,一组维层次 $S=\{l_1,\cdots,l_n\}$,对于某一维层次 l,如果对于每个事实表 F,以及分布聚合函数 af[4],有 $cv(d,F,l,af(m)) = \Pi_{l,af(m)}(\uplus_{i\in 1,\cdots,n}(\Pi_{l,m}(\Gamma^{l}_{l_i} \triangleright \triangleleft cv(d,F,l_i,af(m))))$(其中 $\uplus$ 是对传统并操作的扩展,它并没有消除重复元组),则称在维 d 中层次 l 基于 S 是可汇总的。

* 收稿日期:2006 - 02 - 28
基金项目:国家自然科学基金资助项目(60172012)
作者简介:陆昌辉(1976—),男,博士生。

2 维度汇总性的分类

文献[1,4-7]对一些特殊的维度汇总性问题进行了描述,在此基础上,结合定义 2 的内容,本文将维度汇总性分为不可汇总的、部分可汇总的、完全可汇总的三类。

• 不可汇总的

在定义 2 中,给定一维 d,对于其任一非底层维层次 l,均不存在一个由除其本身之外的非底层维层次组成的集合 S,且 l 基于 S 是可汇总的,则称维 d 是不可汇总的。这又可分为三种情况:第一种情况是由于度量属性本身的原因,使得其不能应用某些分布聚合函数来进行汇总操作;第二种情况是由具体的语义环境而导致了其不可汇总性(比如,快照型度量[1]就不能沿时间维进行汇总操作);第三种情况是由维层次的划分不合理,引发了基础事实数据的重复计算或者遗漏,也就导致了该维是不可汇总的。

• 部分可汇总的

给定某一维 d,并不是其所有非底层维层次 l,均有一个与其对应的由除它本身之外的非底层维层次组成的集合 S,且 l 基于 S 是可汇总的,则称维 d 是部分可汇总的。文献[8]中图 1 描述的 location 维就是部分可汇总的。另外,由于具体的语义环境,也可能会造成维的部分可汇总性,如文献[11]所述。

• 完全可汇总的

给定某一维 d,对于其任一非底层维层次 l,均存在一个与其对应的由除它本身之外的非底层维层次组成的集合 S,且 l 基于 S 是可汇总的,则称维 d 是完全可汇总的。

对于部分可汇总的以及完全可汇总的维,预聚合层次的选取是非常重要的,这往往需要根据具体情况在时间性能以及存储代价中间寻找一个合适的折中点。

3 维度汇总性的判断方法

由语义而导致的不可汇总性,必须根据具体的应用背景来进行判断。在下面描述的判断定理中,本文默认排除了由具体语义而导致不可汇总的情况。

定理 1 在维 d 中,维层次 l 基于维层次集合 S 是可汇总的,当且仅当有下列条件成立:$\Gamma^{l}_{l_{base}} = \biguplus_{l_i \in S} \Pi_{l_{base},l}(\Gamma^{l}_{l^{i}_{base}} \rhd\lhd \Gamma^{l}_{l_i})$

证明 根据定义 2 可知,其判断条件与下述条件等价:

$\Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af^{c}(m)}(\biguplus_{i=1,\cdots,n}\Pi_{l,m}(\Gamma^{l}_{l_i} \rhd\lhd (\Pi_{l_i,m=af(m)}(\Gamma^{l_i}_{l^{i}_{base}} \rhd\lhd f_{base}))))$

// 由于维是由一组相应的维层次来构成的

$\Rightarrow \Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af^{c}(m)}(\biguplus_{i=1,\cdots,n}\Pi_{l,l_i,m=af(m)}(\Gamma^{l}_{l_i} \rhd\lhd (\Gamma^{l_i}_{l^{i}_{base}} \rhd\lhd f_{base})))$

// 由于 af 是分布聚合函数

$\Rightarrow \Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af(m)}(\biguplus_{i=1,\cdots,n}\Pi_{l,m}(\Gamma^{l}_{l_i} \rhd\lhd (\Gamma^{l_i}_{l^{i}_{base}} \rhd\lhd f_{base})))$

// 由于扩展后的投影操作 Π 并没有消除重复元组,故可用 $\Pi_{l_{base},l,m}$ 来替换 $\Pi_{l,m}$

$\Rightarrow \Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af(m)}(\biguplus_{i=1,\cdots,n}\Pi_{l_{base},l,m}(\Gamma^{l}_{l_i} \rhd\lhd (\Gamma^{l_i}_{l^{i}_{base}} \rhd\lhd f_{base})))$

$\Rightarrow \Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af(m)}(\biguplus_{i=1,\cdots,n}(\Pi_{l_{base},l,m}(\Gamma^{l}_{l_i} \rhd\lhd \Gamma^{l_i}_{l^{i}_{base}})) \rhd\lhd f_{base})$

故得出定义 2 中判断维度汇总性的条件与下列条件等价:

$(*)\ \Pi_{l,m=af(m)}(\Gamma^{l}_{l_{base}} \rhd\lhd f_{base}) = \Pi_{l,m=af(m)}(R \rhd\lhd f_{base})$

其中 $R = \biguplus_{i=1,\cdots,n}(\Pi_{l_{base},l}(\Gamma^{l}_{l_i} \rhd\lhd \Gamma^{l_i}_{l^{i}_{base}}))$。

在等式$(*)$中,用 $\Gamma^{l}_{l_{base}}$ 来替换 R 不难得出定理 1 的充分性成立。下面证明其必要性:若 $R \neq \Gamma^{l}_{l_{base}}$,设在维表中存在这样一个元组 $t = (l_{base}, l)$,且该元组在 R 中出现的次数与 $\Gamma^{l}_{l_{base}}$ 在中出现的次数不一样。再令 af 是聚合函数 sum,且 f_{base} 是这样一个事实表,它仅包含唯一的元组$(l_{base}, 1)$,则在$(*)$的左右两边得到的聚合结果是不同的,故存在矛盾。 □

定理2 设 d 为一严格的维[9],其中的维层次 l 基于层次集合 $S=\{l_1,\cdots,l_n\}$ 是可汇总的,当且仅当对于 d 的每个底层维层次成员 $e_{l_{base}}$,在维层次路径集 $\{<l_{base},\cdots,l_1,\cdots,l>,\cdots,<l_{base},\cdots,l_n,\cdots,l>\}$ 中有且仅有一条路径可以上卷到维层次 l 中的相应成员 e_l。

证明 令 $R=\biguplus_{l_i\in S}\Pi_{l_{base}}(\Gamma^{l_i}_{l^i_{base}}\rhd\lhd\Gamma^{l}_{l_i})$,则其可以转换为如下命题的证明:若要有(a)$\Gamma^{l}_{l_{base}}=R$,当且仅当(b)对于 d 的每个底层维层次成员 $e_{\perp_D}$,在维层次路径集 $\{<l_{base},\cdots,l_1,\cdots,l>,\cdots,<l_{base},\cdots,l_n,\cdots,l>\}$ 中有且仅有一条路径可以上卷到维层次 l 中的相应成员 e_l。

- 充分性

假设(a)不成立,则有以下三种情况需要考虑:

(1) 存在元组 $(e_{l_{base}},e_l)\in\Gamma^{l}_{l_{base}}$,且 $(e_{l_{base}},e_l)\notin R$。根据条件(b)可知,在集合 S 中存在唯一的维层次 l_i,$e_{l_{base}}$ 沿着维层次路径 $<l_{base},\cdots,l_i,\cdots,l>$ 上卷到 e_l,故 $\exists y((e_{l_{base}},y)\in\Gamma^{l_i}_{l^i_{base}}\wedge(y,e)\in\Gamma^{l}_{l_i})$,因此 $(e_{l_{base}},e_l)\in\Pi_{l_{base},l}(\Gamma^{l_i}_{l^i_{base}}\rhd\lhd\Gamma^{l}_{l_i})$,从而 $(e_{l_{base}},e_l)\in R$,与 $(e_{l_{base}},e_l)\notin R$ 矛盾。

(2) 存在元组 $(e_{l_{base}},e_l)\in\Gamma^{l}_{l_{base}}$,且该元组在 R 中出现不止一次。

① 则存在一维层次维 $l_i\in S$,有 $(e_{l_{base}},e_i)\in\Gamma^{l_i}_{l^i_{base}}\wedge(e_i,e_l)\in\Gamma^{l}_{l_i}$ 与 $(e_{l_{base}},e_j)\in\Gamma^{l_i}_{l^i_{base}}\wedge(e_j,e_l)\in\Gamma^{l}_{l_i}$ 成立,且 $e_i\neq e_j$,这与 d 严格矛盾。

② 在集合 S 中至少存在两个不同维层次 l_i 与 l_j,且 $(e_{l_{base}},e_l)\in\Pi_{l_{base},l}(\Gamma^{l_i}_{l^i_{base}}\rhd\lhd\Gamma^{l}_{l_i})$,$(e_{l_{base}},e_l)\in\Pi_{l_{base},l}(\Gamma^{l_j}_{l^j_{base}}\rhd\lhd\Gamma^{l}_{l_j})$。故存在 $\exists y((e_{l_{base}},y)\in\Gamma^{l_i}_{l^i_{base}}\wedge(y,e)\in\Gamma^{l}_{l_i})$,$\exists z((e_{l_{base}},z)\in\Gamma^{l_j}_{l^j_{base}}\wedge(z,e)\in\Gamma^{l}_{l_j})$,因此 $e_{l_{base}}$ 既可以沿维层次路径 $<l_{base},\cdots,l_i,\cdots,l>$ 上卷到 e_l,也可以沿路径 $<l_{base},\cdots,l_j,\cdots,l>$ 上卷到 e_l,与(b)矛盾。

(3) 存在元组 $(e_{l_{base}},e_l)\in R$,且 $(e_{l_{base}},e_l)\notin\Gamma^{l}_{l_{base}}$。由于 $(e_{l_{base}},e_l)\in R$,则在集合 S 中存在元素 l_i,有 $\exists y((e_{l_{base}},y)\in\Gamma^{l_i}_{l^i_{base}}\wedge(y,e)\in\Gamma^{l}_{l_i})$ 成立,因此 $(e_{l_{base}},e_l)\in\Gamma^{l}_{l_{base}}$,与 $(e_{l_{base}},e_l)\notin\Gamma^{l}_{l_{base}}$ 矛盾。

- 必要性

假设(b)不成立,则有以下两种情况需要考虑:

(1) 存在元组 $(e_{l_{base}},e_l)\in\Gamma^{l}_{l_{base}}$,且对于集合 S 中的所有元素 l_i,均有 $\neg(\exists y((e_{l_{base}},y)\in\Gamma^{l_i}_{l^i_{base}}\wedge(y,e)\in\Gamma^{l}_{l_i}))$ 成立,则 $(e_{l_{base}},e_l)\notin R$,故与(a)矛盾。

(2) 存在元组 $(e_{l_{base}},e_l)\in\Gamma^{l}_{l_{base}}$,且在集合 S 中至少存在两个不同的元素 l_i 与 l_j,有 $(e_{l_{base}},e_l)\in\Pi_{l_{base},l}(\Gamma^{l_i}_{l^i_{base}}\rhd\lhd\Gamma^{l}_{l_i})$ 与 $(e_{l_{base}},e_l)\in\Pi_{l_{base},l}(\Gamma^{l_i}_{l^i_{base}}\rhd\lhd\Gamma^{l}_{l_j})$ 成立,因此元组 $(e_{l_{base}},e_l)$ 在 R 中至少出现两次,与(a)矛盾。 □

在定理1与定理2的基础上,判断维度汇总性的算法描述如下:

```
Judge _ Dim _ Sum(d)   // d 是要进行判断的维
{
  SumPairs = Φ;  // SumPairs 用来保存二元组对 <l,S>,其中 l 是 d 的维层次,S 是由可上卷到 l
                 的非底层维层次组成的集合,且 l 基于 S 是可汇总的
  Count _ Sum _ Levels = 0;  // Count _ Sum _ Levels 用来记录可汇总的维层次数目
  Counts = 0;  // Counts 用来对维 d 中需要判断其汇总性的维层次进行计数
  For each l_i ∈ d Do   // 对 d 的每个维层次进行扫描
  {
    If (l_i ≠ l_base ∧ l_base ≮ l_i) Then   // l_i 不能是底层维层次,也不能与其存在直接连接
    {
```

```
S = Φ ;  //S 用来保存所有由可上卷到 l_i 的维层次组成的集合,初始化为空
Counts = Counts + 1 ;
For each l_j ∈ d Do
 {
  If (l_j ≠ l_base ∧ l_j ≠ l_i ∧ l_j ≺* l_i) Then   //≺* 为偏序关系的传递闭包
  // 对 d 中上卷到 l_i 的每个非底层维层次进行扫描
     S = S ∪ {l_j} ;  // 将可上卷到 l_i 的非底层维层次添加到集合 S
  }
  For each S_k ⊆ S ∧ S_k ≠ Φ Do
   {
    blRes = JudgeLevelSum(l_i , S_k) ;
    // 根据定理 1 与定理 2 判断 l_i 是否基于 S_k 是可汇总的
    If (blRes = = True) Then
     {
      SumPairs = SumPairs ∪ {< l_i , S_k >} ;  // 对 SumPairs 进行相应更新
      Count _ Sum _ Levels = Count _ Sum _ Levels + 1 ;
      Exit For ;  // 跳出 For 循环
     }
   }
 }
}
If (Count _ Sum _ Levels = = 0) Then
   Return 维 d 是不可汇总的 ;
ElseIf (Count _ Sum _ Levels < Counts) Then
   Return 维 d 是部分可汇总的 ;
Else
   Return 维 d 是完全可汇总的 ;
}
```

在上面描述的维度汇总性判断算法中,同时也为每个可汇总的维层次 l 找到了一个对应的层次集合 S,l 基于 S 是可汇总的。对算法 blRes 为真的部分进行相应修改,可以找到所有满足条件的这样一些集合,为沿着维 d 进行预聚合策略的选取提供了一个初始方案集。

4 维度汇总性问题的处理方法

为了有效地解决这些与维度汇总性有关的问题,在多维建模的概念阶段,可以用如下定义的度量属性聚合模式和维层次聚合模式对那些不可汇总性的情况进行描述:

定义 3(度量属性聚合模式) 多维数据模型 MO 的度量属性聚合模式 MA_{agg} 可以用三元组(m,D_A,agt)集合来表示。其中 m 为该模型的度量属性,D_A 为对该度量属性进行聚合分析的维集合,agt 是由该度量属性允许使用的聚合函数组成的集合。

根据度量属性的实际意义和数据类型,不是每个度量属性都能运用所有的聚合函数而得到正确的聚合结果,这个功能结构视图说明了该模型的度量属性可以运用何种聚合函数沿哪些 D_M 维进行聚合分析。

定义 4(维层次聚合模式) 设 D 为一给定维,l 为其某一维层次,多维数据模型 MO 的维层次聚合模式 DL_{agg} 可以用三元组(l,m,agt)集合来表示。其中 m 为该模型的度量属性,agt 是由对于维层次 l,度量属性 m 允许使用的聚合函数组成的集合。

基于定义3与定义4的描述,在具体实现时,可以采用虚拟度量属性[11]的解决方法,从而拥有了描述维度汇总性的最大灵活性。在该方法中,用度量维的形式来描述真正的度量属性,在事实中除了各个维的主键外,还存在一个称作"数量"的虚拟度量属性,这样就可以将可汇总性定义为维的属性,而维的属性是作为维表的一列或多列存在的,可以对每一个维成员进行相应的属性值设置。因此,根据维成员的自身含义以及用户要求,只需要向各个维(包括度量维)添加相应成员属性,从而就实现了对维度汇总性的各种情况进行灵活约束。该方法在实际工程中得到了有效应用,具体可以参考文献[11]。

5 结论

对于维度汇总性问题,文献[8]利用完整性约束对多维模型中的汇总性进行了研究,文献[10]利用分割约束的形式对异构多维模式下的汇总性推理问题进行了研究,但他们并没有给出具体的解决方案。本文在前人研究的基础上,从维度汇总性的定义出发,将维度汇总性分为不可汇总的、部分可汇总的以及完全可汇总的三类;接着给出了判断维度汇总性的两个相关定理,并对其进行了证明,在此基础上还设计了判断维度可汇总性的算法;最后对维度汇总性问题的相应处理对策进行了介绍。针对那些不可汇总的情况,怎样在保持语义一致性的前提下,通过对维结构的等价变换从而变成可汇总的;针对那些可汇总的情况,怎样选择最佳的汇总策略,这些问题均是今后要研究的重点。

参考文献:

[1] John Horner, Il-Yeol Song, Peter P. Chen. An Analysis of Additivity in OLAP Systems[A]. In Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP[C], Washington, DC, USA, November 12 – 13, 2004:83 – 91.

[2] Lenz H J, Shoshani A. Summarizability in OLAP and Statistical Databases[A]. In Proceedings of the 9th SSDBM Conference[C], Olympia, Washington, USA, 1997:132 – 143.

[3] Shoshani A. OLAP and Statistical Databases: Similarities and Differences[A]. In Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems[C]. Tucson, Arizona, 1997:185 – 196.

[4] Gray J, Bosworth A, Layman A, Pirahesh H. Data Cube: A Relational Operator Generalizing Group-by, Cross-tab and Sub-totals[J]. Data Mining and Knowledge Discovery, 1997, 1(1): 29 – 53.

[5] Holowczak R, Adam N, Artigas J, et al. Data Warehousing in Environmental Digital Libraries[J]. Communications of the ACM, September 2003, 46:172 – 178.

[6] Kim B, Choi K, Kim S, et al. A Taxonomy of Dirty Data[J]. Data Mining and Knowledge Discovery, 2003, 7(1): 81 – 99.

[7] Kimball R, Ross M. The Data Warehouse Toolkit: the Complete Guide to Dimensional Modeling[M]. New York: Join Wiley &Sons, 2002.

[8] Hurtado C A, Gutiérrez C, Mendelzon A. Capturing Summarizability with Integrity Constraints in OLAP[J]. ACM Transactions on Database Systems, 2005, 30(3): 854 – 886.

[9] Pedersen T B. Aspects of Data Modeling and Querying Processing for Complex Multidimensional Data[D]. Ph.D. Thesis, Aalborg University, Denmark, 2000.

[10] Hurtado C A, Mendelzon A O. Reasoning about Summarizability in Heterogeneous Multidimensional Schemas[A]. In Proceedings of the 8th International Conference on Database Theory[C]. London, UK, 2001, 375 – 389.

[11] 林鹏. 战场信息OLAP支撑工具[D]. 长沙:国防科技大学, 2005.