计算机科学 1999Vol. 26№. 6

广义投影和外连接

General Projection and Outerjoin

严哲南　楼荣生

(复旦大学计算机科学系　上海 200433)

Abstract In this paper we introduce mainly the relation between general projection and outerjoin and the relation between general projection and semijoin. We also put forward pseudodistributive law of general projection. Then we provide a more mature theory about general projection.

Keywords General projection, Semijoin, Outerjoin.

在数据仓库中有效地运行聚合查询非常重要，因此加强这方面的理论研究工作变得很重要起来。在文[1]中提出了广义投影的概念，给出了一个初步的查询优化算法。在文[6]中严格定义了广义投影，给出了广义投影的性质，但对广义投影的性质尚不充分，至少没有涉及半连接和外连接与广义投影的关系。半连接和外连接无论在理论上还是在实践上都是很常用的关系运算，因此有必要进一步探讨广义投影与半连接、外连接的关系。本文对文[6]的工作作一补充，讨论了广义投影关于并、伪分配律和广义投影与半连接、外连接的关系。

一、对广义投影的定义及其性质的回顾

在关系运算过程中有时允许多重关系，设 A_1，A_2，…，A_n 是 n 个多重集定义 $A_1\times A_2\times\cdots\times A_n=(A_1\times A_2)\times\cdots)\times A_n$，则 $A_1\times A_2\times\cdots\times A_n$ 的子集 r 为 $A_1,A_2,\cdots,A_n$ 中的 n 元多重关系。(注：多重集用 $\{\}^*$ 表示，例如 $\{a,a,a\}^*$，就是一个多重集，它也可表示为压缩形式$(a,3)$，表示 a 在集合中重复了三次)。

在本文中，若不特别说明，我们用小写字母 r,s 表示关系或多重关系，用大写字母 R,S 表示关系模式，若不发生混淆，也用 R 表示 r 的属性集。

多重关系的合并定义：$r+s=s+r=\sum(c,z)$ (其中(c,z)满足$(c,z)\subseteq r$ 且 $c\notin s$，或 $c\notin r$ 且$(c,z)\subseteq s$，或$(c,x)\subseteq r$ 且$(c,y)\subseteq s$ 且 $z=x+y$)。如$\{a,a,a,b,b\}^*=\{a,a,a\}^*+\{b,b\}^*=(a,3)+(b,2)$。

我们沿用文[6]中的定义和符号：$\boldsymbol{R}$ 是关系模式的属性集，$\forall\ i,1\leqslant i\leqslant n,A_i\in\boldsymbol{R}$；$\forall\ j,1\leqslant j\leqslant m,B_j\in\boldsymbol{R}$；约定用粗体字表示一个序列 $\boldsymbol{G}=A_1A_2\cdots A_n,\boldsymbol{B}=B_1B_2\cdots B_m$，$\forall\ i,1\leqslant i\leqslant m,H_i\in\{min,max,sum,count,avg\}$，$\boldsymbol{H}(\boldsymbol{B})=H_1(B_1)H_2(B_2)\cdots H_m(B_m)$，$\forall\ k,1\leqslant k\leqslant l,\boldsymbol{R}_k$ 是属性集，$\boldsymbol{R}_k\subseteq\boldsymbol{R}$，记 $F_1(\boldsymbol{R}_1)F_2(\boldsymbol{R}_2)\cdots F_l(\boldsymbol{R}_l)$为 $\boldsymbol{F}(\boldsymbol{R})$，$\boldsymbol{H}(\boldsymbol{B}=\boldsymbol{F}(\boldsymbol{R}))$为 $\boldsymbol{HF}(\boldsymbol{R})$。因为 avg ()可由 sum()和 $count$()得到，因此在本文的讨论中，侧重讨论其它四种聚合运算。若不特别指出聚合运算 H 的范围，则 $H\in\{min,max,sum,count\}$，下面三个定义是文[6]所用定义：

①$\pi^1_{\boldsymbol{G},\boldsymbol{H}(\boldsymbol{B})}(r)=\{\boldsymbol{ah}\mid\boldsymbol{a}\in\Pi_{\boldsymbol{G}}(r),\boldsymbol{h}=\boldsymbol{H}(\Pi^d_{\boldsymbol{B}}\sigma_{\boldsymbol{G}=\boldsymbol{a}}(r))\}$，称为第一类广义投影，简记为 $\pi^1_{\boldsymbol{G},\boldsymbol{H}}(r)$，若不混淆，也可简单记 π^1 为 π。②称 $\pi^2_{\boldsymbol{G},\boldsymbol{F}(\boldsymbol{R})}(r)$为第二类广义投影，简写为 $\pi^2_{\boldsymbol{G},\boldsymbol{F}}(r)$，$\pi^2_{\boldsymbol{G},\boldsymbol{F}}(r)=\{\boldsymbol{af}\mid t\in r,\boldsymbol{a}=\Pi_{\boldsymbol{G}}(t),\boldsymbol{f}=\boldsymbol{F}(t)\}^*$，其中 $\boldsymbol{F}$ 表达成 SQL，实现时不必用 Groupby 子句。③$\pi^3_{\boldsymbol{G},\boldsymbol{H}(\boldsymbol{F}(\boldsymbol{R}))}(r)=\{\boldsymbol{ah}\mid\boldsymbol{a}\in\Pi_{\boldsymbol{G}}(r),\boldsymbol{h}=\boldsymbol{H}(\Pi^d_{\boldsymbol{B}}\sigma_{\boldsymbol{G}=\boldsymbol{a}}(\pi^2_{\boldsymbol{G},\boldsymbol{B}=\boldsymbol{F}(\boldsymbol{R})}(r))$，称为第三类广义投影，简写为 $\pi^3_{\boldsymbol{G},\boldsymbol{H}(\boldsymbol{F})}(r)$。

本文中还使用如下符号：$j=1,2,\boldsymbol{G}^j\subseteq\boldsymbol{R}^j,\boldsymbol{B}^j\subseteq\boldsymbol{R}^j,\boldsymbol{I}=\boldsymbol{R}^1\cap\boldsymbol{R}^2$，对 $\boldsymbol{G}^1,\boldsymbol{G}^2$ 也是如此。$\boldsymbol{H}^j=H^j_1H^j_2\cdots H^j_{m_j},\forall\ H^j_i\in\{min,max,sum,count\}$。$L^j=L^j_1\cdots L^j_{m_j},F^j=f^j_1f^j_2\cdots f^j_{m_j}$，当 $H^j_i=sum$ 时，$f^j_i=L^j_i*X_j$；当 $H^j_i\in\{min,max\}$时，$f^j_i=L^j_i$。除非另外定义，以下相同符号与此处意义相同。

在文[6]中还提出了投影划分的定义及性质：设 $G\subseteq R$，定义关系(或多重关系)r 上的二元关系 E：$E=\{(a,b)\mid a\in r,b\in r,a[G]=b[G]\}$。可以验证 E 是 r 上的等价关系(因为它是自反的，对称的，传递

的)。E 确定 r 上的一个划分 $D_1,D_2,\cdots,D_k$,称划分 $D_1,D_2,\cdots,D_k$ 为 G 确定的 r 上的投影划分,记为 $r/G=\{D_1,D_2,\cdots,D_k\}$。设 D_i 中的任何元组的 d_i 都是它的代表元为 d_i,记 $D_i=[d_i]$。很清楚,$D_1[G],D_2[G],\cdots,D_k[G]$ 是 $r[G]$ 的划分。

在文[6]中还给出了广义投影的许多性质,我们可以不加证明地适当推广到多重关系,并可以总结为以下几个公式:

公式 1 $\Pi_G(r_1),\cdots,\Pi_G(r_n)$ 两两互不相交,则 $\pi_{G,H(B)}(\sum_{i=1}^{n} r_i)=\bigcup_{i=1}^{n}\pi_{G,H(B)}(r_i)$。

公式 2 $G_1\subseteq G_2$,$s\geqslant m$,若 $s\geqslant i>m$,$H'_i\in\{min,max,sum,count,avg\}$,若 $1\leqslant i\leqslant m$,$H'_i=H_i$ 在 $H_i\in\{min,max,sum\}$ 时成立,$H_i=count$ 当 $H'_i=sum$ 时成立,则:

$$\pi_{G_1,H_1(B_1)H_2(B_2)\cdots H_m(B_m)}(r)=\pi_{G_1,H'_1(Q_1)H'_2(Q_2)\cdots H'_m(Q_m)}(\pi_{G_2,Q_1=H_1(B_1)Q_2=H_2(B_2)\cdots Q_s=H_s(B_s)}(r))$$

公式 3

1) $\pi^1_{G,H(B)}\sigma_{p(G)}=\sigma_{p(G)}\pi^1_{GH(B)}$

2) $\pi^2_{G,F}\sigma_{p(G)}=\sigma_{p(G)}\pi^2_{G,F}$

3) $\pi^2_{G,H(F)}\sigma_{p(G)}=\sigma_{p(G)}\pi^3_{G,H(F)}$

公式 4

$$\pi_{G^1,G^2,H^1(B^1),H^2(B^2),X=count(*)}(r^1\times r^2)=\pi^2_{G^1,G^2,F^1,F^2X=X^1\cdot X^2}(\pi_{G^1,L^1=H^1(B^1),X_2=count(*)}(r^1)\times\pi_{G^2,L^2=H^2(B^2),X_1=count(*)}(r^2))$$

公式 5 设 r^1、r^2 是关系,$G^1=G\cap R^1$,$G^2=G\cap R^2$,$G\subseteq R^1\cup R^2$,则当 $I\subseteq G$ 时,1)成立,否则在一般情况下,2)成立。

1) $\pi_{G,H^1(B^1),H^2(B^2)}(r^1\infty r^2)=$

$$\pi^2_{G,F^1,F^2}(\pi_{G^1,L^1=H^1(B^1),X_2=count(*)}(r^1)\infty\pi_{G^2,L^2=H^2(B^2),X_1=count(*)}(r^2))$$

2) $\pi_{G,H^1(B^1),H^2(B^2)}(r^1\infty r^2)=$

$$\pi^3_{G,H^1(F^1),H^2(F^2)}(\pi_{G^1,L^1=H^1(B^1),X_2=count(*)}(r^1)\infty\pi_{G^2,L^2=H^2(B^2),X_1=count(*)}(r^2))$$

本文将利用这些公式演绎出有关广义投影与半连接、外连接关系的定理。

二、广义投影的性质

在文[6]中给出了广义投影关于关系并、差的分配律,是比较特殊的情况,没有对多重关系作讨论,以下定理 1 将给出广义投影关于多重关系并、差的伪分配律,比文[6]中的讨论更适用一般情况。

2.1 伪分配律

引理 1 r^1 和 r^2 是多重关系,则对任意 $g\in r^1+r^2$,$[g]=[g]_1+[g]_2$。其中 $[g]\in(r^1+r^2)/G$;若 $g\in r^i$,$[g]_i\in r^i/G$,否则 $[g]_i=\varnothing$。

证明:(略)。

定理 1(伪分配律) r^1 和 r^2 是多重关系,$H(B)=H_1(B_1)H_2(B_2)\cdots H_m(B_m)$,对 $\forall i$,$1\leqslant i\leqslant m$,$H_i\in\{min,max,sum\}$,$B=B_1B_2\cdots B_m$,则:

$$\pi_{G,H(B),X=count(*)}(r^1+r^2)=\pi^3_{G,H(Q),X=X1+X2}(\pi_{G,Q=H(B),X1=count(*)}(r^1)+\pi_{G,Q=H(B),X2=count(*)}(r^2))$$

证明:(略)

推论 1 r,s 是关系,$H(B)=H_1(B_1)H_2(B_2)\cdots H_m(B_m)$,对 $\forall i$,$1\leqslant i\leqslant m$,$H_i\in\{min,max\}$,$B=B_1B_2\cdots B_m$,则

$$\pi_{G,H(B)}(r\bigcup s)=\pi_{G,H(Q)}(\pi_{G,Q=H(B)}(r)\bigcup\pi_{G,Q=H(B)(S)})$$

证明:(略)。

2.2 广义投影与半连接的关系

半连接在分布式系统中很重要,因此对它的优化的研究就很重要,这里解决了有关广义投影与半连接的关系问题,必然对半连接的优化起到重要的作用。半连接(用符号∞表示)的定义为 $r\infty s=\Pi_R(r\infty s)$。

$G\subseteq R'$,$B\subseteq R^1$,r^1 是关系,r^2 是多重关系。

定理 2 $I\subseteq G$,$\forall H_i\in\{min,max,sum,avg,count\}$,则 $\pi_{G,H(B)}(r^1\infty r^2)=\pi_{G,H(B)}(r^1)\infty r^2$

证明:(略)。

推论 2 $H'=H'_1H'_2\cdots H'_m$,对 $\forall i$,$H'_i=H_i\in\{min,max,sum\}$ 或 $H'_i=sum$ 且 $H_i=count$,则在一般情况下,即不必要求 $I\subseteq G$,此时 $\pi_{GH(B)}(r^1\infty r^2)=\pi_{GH'(L)}(\pi_{GI,L=H(B)}(r^1\infty r^2)$。

2.3 广义投影与外连接的关系

外连接是 SQL 中的重要运算,它与等值连接的不同之处在于前者在连接过程中把悬浮元组与和它相连接的关系的空元组连接。在文[7]中给出了外连接(用符合▷表示)的等价定义:$r\triangleright s=r\infty s\cup r\sim s\times\varnothing[s]$,$r\sim s=r-r\infty s$。

定理 3 $\forall H'_i\in\{min,max,sum\}$,$r^1$,$r^2$ 是非多重关系,有:

1) $R^1\cdot I\subseteq G^1$,$R^2\cdot I\subseteq G^2$,则

$$\pi_{G^1,G^2,H^1(B^1),H^2(B^2)}(r^1\triangleright r^2)=\pi^2_{G^1,G^2,F^1,F^2}(\pi_{G^1,L^1=H^1(B^1),X_2=count(*)}(r^1)\triangleright\pi_{G^2,L^2=H^2(B^2),X_1=count(*)}(r^2))$$

2) $\pi_{G^1,G^2,H^1(B^1),H^2(B^2)}(r^1\triangleright r^2)=\pi^3_{G^1,G^2,H^1(F^1),H^2(F^2)}(\pi_{G^1,L^1=H^1(B^1),X_2=count(*)}(r^1)\triangleright$

$\pi_{G^2,L^2=H^2(B^2),X_1=count(*)}(r^2))$

若 $\pi_{G^1L^1=H^1(B^1),X_2=count(*)}(r^1)\triangleright$ $\pi_{G^2L^2=H^2(B^2),X_1=count(*)}(r^2)$的属性 X_1 对应的值为空值(请注意这里所说的空值不是指 0,而是外连接所生成的空值),则令属性 X_1 取值为 1。

证明:(略)

例 $R_1(A^1B^1C^1),R_2(A^2B^2C^2D),A^1=A^2$,则:

$\pi_{A^1A^2,sum(B^1),sum(B^2)}(r^1\triangleright r^2)$

$=\pi^2_{A^1A^2,L^1*X_1,L^2*X_2}(\pi_{A^1L^1=sum(B^1)X_2=count(*)}(r^1)\triangleright$ $\pi_{A^2L^2=sum(B^2)X_1=count(*)}(r^2))$

$\pi_{A^1A^2D,sum(B^1),max(B^2)}(r^1\triangleright r^2)$

$=\pi^3_{A^1A^2D,sum(L^1*X_1),max(L^2),X=X_1*X_2}$

$(\pi_{A^1L^1=sum(B^1)X_2=count(*)}(r^1)\triangleright\pi_{A^2L^2=max(B^2)X_1=count(*)}(r^2))$

注意,如果属性 X_1 对应的值为空值,则令属性 X_1 取值为 1。

应用定理 3 可以在两方面优化查询操作:第一、若关系在主键属性上的值冗余比较大,利用该公式有可能减少外连接的计算量。第二、如果等式右边的广义投影已有相应的固化视图则可以直接利用固化视图计算,避免重复计算。

推论 3 $\forall H_i^j\in\{min,max\}$,是非多重关系,有:

1)若 $R^1.I\subseteq G^1,R^2.I\subseteq G^2$,则

$\pi_{G^1,G^2,H^1(B^1),H^2(B^2)}(r^1\triangleright r^2)=\pi_{G^1,H^1(B^1)}(r^1)\triangleright$ $\pi_{G^2,H^2(B^2)}(r^2)$

2)$\pi_{G^1,G^2,H^1(B^1),H^2(B^2)}(r^1\triangleright r^2)=\pi_{G^1,G^2,H^1(L^1),H^2(L^2)}$ $(\pi_{G^1,L^1=H^1(B^1)}(r^1)\triangleright\pi_{G^2,L^2=H^2(B^2)}(r^2))$

推论 4 $\forall H_i^j\in\{\min,\max\}$,$I\subseteq G_1$,则 $\pi_{G^1,H^1(B^1)}(r^1\triangleright r^2)=\pi_{G^1,H^1(B^1)}(r^1)\triangleright r^2$

结论 结合前面公式 1～5 和本文的三个定理,我们可以把广义投影的性质总结为 8 个公式:

公式 6 定理 1。

公式 7 定理 2。

公式 8 定理 3。

作为对比,一般非广义投影(用符号 Π 表示)和复本投影(用符号 Π^d 表示)有如下性质:

公式 9:$\Pi\Pi=\Pi,\Pi^d\Pi^d=\Pi^d$

公式 10 1)$\Pi(r\cup s)=\Pi(r)\cup\Pi(s)$,

$\Pi^d(r+s)=\Pi^d(r)+\Pi^d(s)$;

2)$\Pi(r-s)\supseteq\Pi(r)-\Pi(s)$

公式 11 $Y\subseteq X$,则 $\Pi_Y=\Pi_Y\Pi_X$,$\Pi^d_Y=\Pi^d_Y\Pi^d_X$

公式 12 $\sigma_{p(X)}\Pi_X=\Pi_X\sigma_{p(X)}$,$\sigma_{p(X)}\Pi^d_X=\Pi^d_X\sigma_{p(X)}$。其中($p(X)$)是仅有 X 中的属性做变元的条件表达式。

公式 13 $X\subseteq R,Y\subseteq S,\Pi_{X+Y}(r\times s)=\Pi_X(r)\times\Pi_Y(s)$

公式 14 $R\cap S\subseteq X\subseteq R\cup S,X_1=X\cap R,X_2=X\cap S$,则

$\Pi_X(r\infty s)=\Pi_{X_1}(r)\infty\Pi_{X_2}(s)$,$\Pi_X(r\infty s)=\Pi_{X_1}(r)\infty\Pi_{X_2}(s)$

通过对比和归纳可以看出广义投影是投影概念的推广,但它的性质与投影的性质不同,它需要满足一定的条件才能成立关系和差的分配律,并且需要特定的条件才能与选择操作,笛卡儿积,连接,半连接以及外连接交换。

参考文献

1 Gupta A, et al. Aggregate-Query Processing in Data Warehouse Environments. In: Proceeding of the 21st VLDB conference. Zurich, Switzerland, 1995

2 Quass D, Widom J. On-Line Warehouse View Maintenance. ACM SIGMOD, 1997

3 Zhuge Y, et al. Multiple View Consistency for Data Warehousing

4 Zhuge Y, et al. View maintenance in a warehousing environment. SIGMOD, May 1995: 316～327

5 Zhuge Y, et al. Consistency algorithms for multi-source warehouse view maintenance

6 严哲南,楼荣生,范皓. 广义投影及其性质. 计算机科学,待发表

7 楼荣生. SQL 外连接及其性质. 广西师范大学学报, 15(1)

* r∞s 在这里的含义是在 r 与 s 所有公共属性上相等的等值连接,而不是自然连接。