计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
43(2):288~294, 2006

基于直方图的 XPath 含值谓词路径选择性代价估计

王 宇[1] 孟小峰[2] 王 珊[2]
[1](河北大学计算中心 保定 071002)
[2](中国人民大学信息学院 北京 100872)
(sandpiperwy@yahoo.com.cn)

Using Histograms to Estimate the Selectivity of XPath Expression with Value Predicates

Wang Yu[1], Meng Xiaofeng[2], and Wang Shan[2]
[1](*Computer Center*, *Hebei University*, *Baoding* 071002)
[2](*Information School*, *Renmin University of China*, *Beijing* 100872)

Abstract Selectivity estimation of path expressions is the basis of XML query optimization and also intense research interest. A path expression may contain multiple branches with value predicates. Some of the values and the nodes of the XML data are highly correlated. Previous methods of selectivity estimation rarely take that relation into consideration, and assume, instead, that the selectivity of attribute values on different nodes and structures is independent and uniform. In this paper, a novel value histogram is proposed, which captures the correlation between the structures and the values in the XML data. Also defined are six operations on the value histograms as well as on the traditional histograms that capture nodes positional distribution. Based on such operations, the selectivity of any node (or branch) in a path expression can be estimated. Experimental results indicate that the method provides accuracy especially in cases where the distribution of the value or structure of the data exhibit a certain correlation without any independent assumption.

Key words XML; query optimization; selectivity; histogram; predicate

摘 要 路径选择性代价估计是 XML 查询优化的基础,也是研究的热点. 目前的方法采用大量正态分布和独立性分布假设是造成误差的根本原因. 定义了一种新颖的值-位置直方图用于统计 XML 数据中的结构和值的分布情况,并提出了 6 种直方图运算. 在此基础上,给出用直方图计算估计路径中任一结点选择性的方法. 实验证明,这种方法无需独立性分布假设,也能在数据结构和数值分布不均匀的情况下,精确地估计路径选择性代价.

关键词 XML;查询优化;选择性;直方图;谓词

中图法分类号 TP311.13

1 引 言

用 XPath 表示的多谓词复杂路径是 XQuery 的核心表达式,也是影响 XML 查询执行效率的关键因素. 如何优化执行多谓词复杂路径是人们关注的焦点问题. 复杂路径表达式包含多个谓词分支,如何精确地估计位于不同分支结点的选择性成为研究

收稿日期:2004-10-09;修回日期:2005-06-22
基金项目:国家自然科学基金项目(60073014,60273018);国家"八六三"高技术研究发展计划基金项目(2002AA116030);教育部科学技术重点基金项目(03044);河北大学博士科研启动基金项目(Y2005050)

的热点.

在含值谓词复杂路径中,既隐含数据之间的嵌套结构,更有对分散在结构中的值的计算. 为了精确地计算结点的选择性,需要综合考虑表达式中所有结点对该结点的影响. 传统方法在计算时需使用正态分布和独立性分布等假设. XML 数据有复杂的层次结构,相关结点的分布以及结点值的分布很难满足传统代价计算常用的分布假设,导致代价估计的误差.

我们设计了一种新颖的二维直方图,正确地反映 XML 数据中值与结构的关系,提出了一种基于直方图计算的复杂路径选择性代价计算方法,把值与结构的相关性隐含在直方图的计算中. 构造了模式与直方图相结合的统计信息模型,给出了利用模式信息生成高效代价计算树的方法. 通过与 XSketch 方法的比较实验,从存储代价、计算效率和计算精确性 3 个方面证明了本文提出方法的有效性和先进性.

2　相关工作比较

目前 XML 信息统计和代价计算方法主要分为两种:一种基于模式,从宏观角度掌握数据的整体分布情况,基于一定的分布假设计算复杂路径的查询代价[1~5];另一种用直方图存储数据之间的位置关系,计算路径或者结构连接的代价[6,7].

Lore[1]方法从数据中提取模式信息(DataGuide),基于父子结点分布的独立性假设计算长路径的选择性. 其计算方法基于 Markov 链思想,只适用于没有分支条件的简单路径. 文献[2]采用 Markov 表保存路径信息,相当于 Lore 方法的改进和丰富. 文献[3]等采用后缀树方法保存结构信息,需将查询分解为简单的子查询分别计算,然后按照独立性假设组合. 不支持数值等原子类型,只能够处理等值谓词.

二维位置直方图(PH)[6]或一维位置直方图(PLH)[7]方法可用于计算结构连接运算代价. 位置直方图只保存数据的位置关系,不能直接用于计算含值路径的选择性. 本文在计算中引用了上述结构,但计算方法有本质的不同. 并且,本文提出的 PD 和 PA 运算解决了用直方图计算父子关系代价的问题.

已知的综合利用模式信息和直方图的工作有两个,StatiX[5]和 XSketch[4,8,9]. StatiX[5]用一维直方图方法统计分支结点的路径分布信息,仅能根据父亲计算儿子的分布情况. XSketch[4,8,9]利用模式信息统计数据分布的整体情况,采用传统多维直方图保存多个值对某类结点的影响. 但这种方法需要预先分析数据之间的相关性. 当路径中包含多个含值谓词时,仍然需要独立性分布假设综合计算结果.

3　值-位置直方图

XML 数据通常被模型为树状结构,XML 树中的每个结点可对应惟一的一个编码($start$, end),称为位置标识. 位置标识满足如下特点:①祖先位置标识严格地包含后代位置标识;②兄弟位置标识互不重叠. 根结点的位置标识区间最大,我们称之数据编码区间,如图 1 所示:

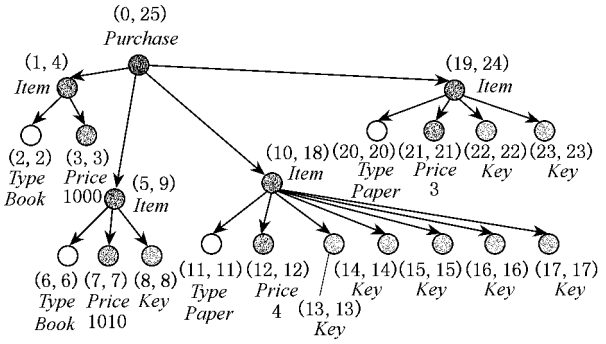

Fig. 1　Example data tree and its region codes.

图 1　实例数据与位置标识

设 XML 数据编码区间为[min, max],某类型结点内容的值域为[V_{min}, V_{max}],在平面坐标系中,x 轴为值所属结点的编码区间,y 轴为值域. 某类结点集合在坐标系上的分布称为值-位置分布图.

图 1 中 $Price$ 的值-位置分布如图 2(a)所示. 点 a 的坐标为(7,1010),表明其位置标识的 $Start$ 为 7,$Price$ 值为 1010.

定义 1. 将值-位置分布图的 x 轴等分为 g 个格,y 轴根据不同类型和值域范围划分为 m 格,分别统计结点在每个格中的结点个数,构成值-位置直方图,记为 $\boldsymbol{VH}$.

值-位置直方图的每一格统计的是结点位置标识中 $Start$ 和结点的值包含在本格区间中的结点个数,如图 2(d)所示. $\boldsymbol{VH}[1][3]=1$ 表示 $Price$ 值在[757.5,1010], $Start$ 值在[6.25,12.5]区域范围内的结点个数为 1 个.

值-位置直方图体现值的分布与结构之间的关系. 若有谓词约束某值域,可以通过值-位置直方图很容易地得到满足谓词的结点位置标识的 $Start$ 分

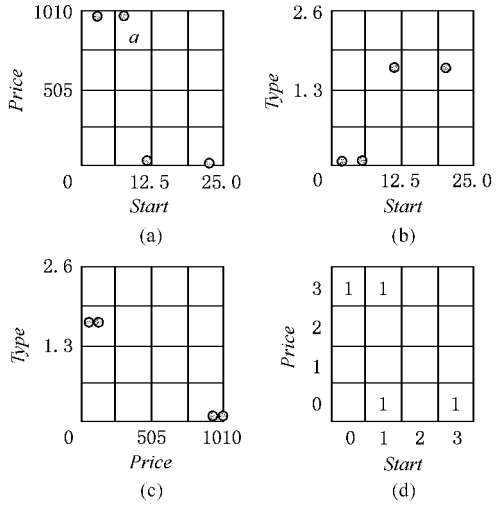

Fig. 2 *VH* vs. two dim. value distribution. (a) Value/code distribution of *Price*; (b) Value/code distribution of *Type*; (c) Value distribution of *Type* and *Price*; and (d) ***VH*** of *Price*.

图 2 值-位置直方图与二维值分布比较.(a) *Price* 值-位置分布;(b) *Type* 值-位置分布;(c) 二维值分布;(d) *Price* 值-位置直方图

布情况. 而 *Start* 值对应结点在树中的先序遍历值,隐含结点的位置分布情况. 因此,根据值-位置直方图可以获得满足谓词的结点的位置分布. 如查询(例 1):/purchase/item[*Type* =" book "][*Price* > 500]/*Key* 中求 *Item* 的选择性. 通过 *Price* 和 *Type* 的值-位置分布(图 2)可以看出,满足谓词 *Price* > 500 与 *Type* =" *book* "的 *Start* 值范围重合,均在[0, 12.5]内. 这个结果还可用于进一步计算 *Key* 的选择性. 而如果采用 XSketch[8]的方法,用二维值直方图反映不同 *Price* 和 *Type* 值域内 *Item* 的个数,如图 2(c)所示. 也可得到 *Item* 的个数为 2,但 *Item* 的位置是未知的,只能根据独立性假设进一步计算 *Item* 对 *Key* 的影响. 如果相关结点大于两个,则需要更高维的直方图,存储和计算代价会成幂级增长甚至导致方法的不可行. 这实际上只是孤立计算方法的一种改进. 而值-位置直方图分别统计不同值的位置分布,利用计算组合分布情况,真正建立了值与结构之间的桥梁,既保证的存储和计算效率,又不会丢失相关性信息.

体现结构分布的位置直方图有两种,Wu 等人[6]提出的 PH 和 Wang 等人[7]提出的 PLH. 后面我们将以 PH 为代表进行论述. 图 3 是一个 PH 实例图. 图 3(a)是图 1 中除 *Key* 外其他结点的二维分布图,图 3(b)为 *Price* 类结点的 4×4 ***PH***. 关于位置直方图的详细介绍请参考具体文献.

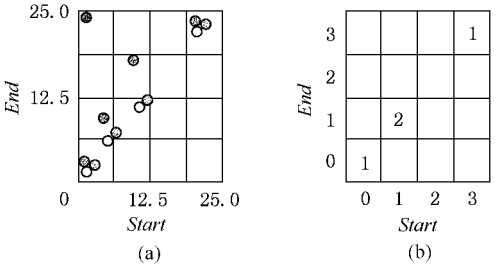

Fig. 3 An example of PH. (a) Distribution graph and (b) ***PH*** of *Price*.

图 3 位置直方图实例.(a) 二维数据分布;(b) Price 位置直方图

4 直方图基本运算

直方图运算有 6 种:值选始位置(V)和始位置转换(S)用于谓词计算,选后代(D)和选祖先(A)用于结构嵌套计算,自除后代(PD)和自除祖先(PA)用于计算父子关系时去掉祖先或后代. 参与运算的是直方图,运算的结果也是直方图. 考虑到篇幅问题,我们以 ***PH*** 为参数简要说明运算. PLH 与其思想相同,但计算方法稍有差别.

值选始位置运算(V):在 ***VH*** 不同列中分别统计满足值域条件的结点个数,构成一维始位置直方图 ***SH***.

始位置转换运算(S):统计 ***PH*** 每列结点之和,与 ***SH*** 对应格之比作为谓词的选择性;乘以 ***PH*** 对应列的格中结点个数,得到满足谓词的位置直方图 $\boldsymbol{PH}^P$.

选后代运算(D):对后代位置直方图中每个格,根据其祖先位置直方图 $\boldsymbol{PH}^A$ 对应区域的格中结点个数,确定其选择性 $SelPH[i][j]$. 根据选择性,得到满足祖先的后代位置直方图.

选祖先运算(A):对祖先位置直方图中每个格,根据其后代位置直方图 $\boldsymbol{PH}^D$ 对应区域的格中结点个数,确定其选择性 $SelPH[i][j]$. 根据选择性,得到满足后代的祖先位置直方图.

自除后代运算(PD):在位置直方图中去除嵌套的后代结点.

自除祖先运算(PA):在位置直方图中去除嵌套祖先结点.

6 种运算的计算复杂度随直方图格数变化. 设 ***VH*** 格数为 $m \times g$,***PH*** 格数为 $g \times g$,则 V 运算复杂度为 $O(m \times g)$,S 运算复杂度为 $O(g^2)$,A,D,PA,PD 运算均为两个二维直方图的嵌套迭代,算法复杂度为 $O(g^4)$.

5　路径选择性计算(PM)

计算某结点在路径中的选择性需构造代价树(CT).后根序遍历执行 CT 得到的位置直方图即反映满足路径约束的结点的位置分布情况.在本节中,首先介绍简单谓词和简单路径中结点的选择性计算,然后通过实例分析多谓词复杂路径表达式中结点的选择性计算.

5.1　简单谓词选择性计算

我们称形如[$Type$="book"]的谓词表达式为简单谓词.计算简单谓词对结点的选择性的方法为以结点的 VH 为参数进行 V 运算,得到始位置直方图 SH.与结点 PH 进行 S 运算,得到满足谓词的位置直方图 PH^P.

5.2　简单路径选择性计算

若路径表达式为 a/b 或者 $a//b$,其中 a,b 为结点名,则该路径表达式为简单路径.简单路径中没有谓词出现.

若表达式为 $a//b$,通过 A 运算或 D 运算即可获得满足路径的 a 结点或 b 结点的 PH.若路径表达式为 a/b,需要在相应运算之前,先进行 PA 或 PD 运算以去掉祖先或者后代结点.

简单路径和简单谓词是两种基本的代价计算.所有复杂路径表达式均可分解为简单路径和简单谓词.

5.3　多谓词复杂路径选择性计算

多谓词复杂路径形如:$n_1[p_1^1][p_1^2]\cdots[p_1^{k1}]\phi n_2[p_2^1][p_2^2]\cdots[p_2^{k2}]\phi\cdots\phi n_m[p_m^1][p_m^2]\cdots[p_m^{km}]$,其中 n 为结点名,p 为谓词路径,ϕ 为"/"或"//".估计路径中某结点 n_x 的选择性,需从 3 方面计算:祖先对其的选择性,后代对其的选择性和谓词对其的选择性.下面分别论述.

计算来自祖先的选择性需从路径中的最左结点开始沿路径做 D 运算,如果结点之间的关系为父子,还应在子结点参加运算之前做 PD 运算.如果某结点有谓词约束,则应首先计算谓词约束对该结点的选择性,利用满足谓词约束的位置直方图参加运算.例如,若计算路径(例 2)$n_1/n_2//n_3//n_4[a>v]$ 中祖先结点对 n_3 的选择性,首先对 $PH(n_2)$ 做 PD 运算,然后与 $PH(n_1)$ 做 D 运算,再与 $PH(n_3)$ 做 D 运算.

计算来自后代的选择性则相反,从路径中最右结点开始沿路径做 A 运算.若结点之间的关系为父子,还需加入 PA 和 PD 运算.若某后代结点有谓词约束,则也应首先计算谓词约束对该后代结点的选择性.如计算例 2 中后代结点对 n_3 的选择性,首先计算谓词对 n_4 的选择性,利用满足谓词的 n_4 的 PH 沿路径做 A 运算,计算 n_4 对 n_3 的选择性,如图 4(a)所示:

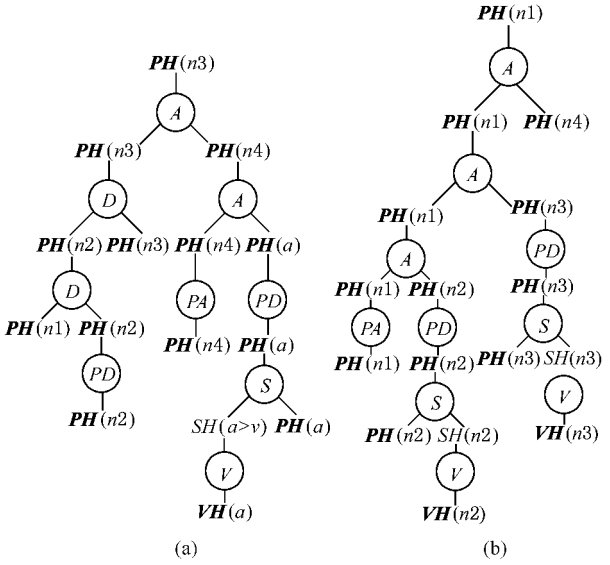

Fig. 4　An example CT. (a) $n_1/n_2//\underline{n_3}//n_4[a>v]$ and (b) $\underline{n_1}[n_2>v][n_3>w]//n_4$.

图 4　CT 实例.(a) $n_1/n_2//\underline{n_3}//n_4[a>v]$;(b) $\underline{n_1}[n_2>v][n_3>w]//n_4$

计算谓词对某结点的选择性与计算后代对其的选择性相似.若存在多个谓词,求分支结点的选择性是关键问题,需要按顺序分别计算不同谓词的选择性.如(例 3):$n_1[n_2>v][n_3>w]//n_4$ 中,n_1 为分支结点,其选择性 CT 如图 4(b)所示.首先计算谓词 n_2 对其的选择性,然后计算谓词 n_3,最后计算后代 n_4 的选择性.

PM 方法的中心思想是利用一系列直方图运算,综合路径中所有结点对估计结点选择性的影响.路径中结点间结构和值的相关性反映在直方图的不同区域计数的变化中.因此 PM 方法无需事先对结点的相关性进行分析,也无需任何独立性分布假设.PM 方法既可用于计算路径中任一点在整个路径中的选择性,也可用于计算任一点在部分路径中的选择性.但是,这种方法的缺点在于运算次数多,导致效率的下降.如例 3 中结点个数为 4,但 CT 中的运算却高达 10 个.为了提高代价计算的效率,我们采用一种改进的方法——模式与直方图相结合的方法(SGM).

6 模式指导的路径选择性计算(SGM)

6.1 统计信息模型

SGM 方法计算复杂路径选择性所需的统计信息由 3 部分组成:模式信息、不同类型结点的位置直方图和含值结点的值-位置直方图.

定义 2. 统计信息模型(*SI*)用三元组[N,E,T]表示. N 为结点类型集合,$\forall n \in N$,有 $tagName$(Eid,C,R,V,L),$tagName$ 为结点名称,Eid 惟一地标识该结点;C 为该类型结点对应数据结点的个数;R 为递归嵌套标记,标记结点是否在递归嵌套的圈中;V 为其内容值域范围,若 n 不含值则 V 为空;L 为指向直方图的指针,每个结点对应一个位置直方图,如果该结点含值,则还有值-位置直方图. E 为边集合,表示结点之间的嵌套关系,为了简单起见,不区分属性边. T 为直方图与 Eid 对照表. 为了方便查找,我们以 Eid 为关键字在表上建立了 $B+$ 树索引.

我们采用绝对路径方法生成模式图[4]. 在 XML 树中同名的结点如果入边路径不同,则在模式图中视为不同类型的结点. 图 5 为统计信息模型实例:

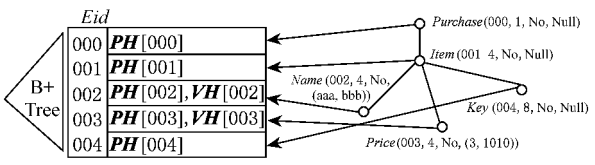

Fig. 5 An example of statistic information model.

图 5 统计信息实例

影响统计信息存储代价的因素有 4 个:模式的结点个数(n);含值结点个数(nv);直方图格数(g,m)和存储方法. 数据模式的结点个数决定了位置直方图的个数,含值结点个数决定了值-位置直方图的个数. 直方图格数决定每个直方图的规模. PH+VH 统计信息存储代价为 $n + g^2 \times n + g \times m \times nv$. PLH+VH 统计信息存储代价为 $n + g \times 3 \times n + g \times m \times nv$. 两种存储总代价均随格数呈幂级增长. 当格数很大时,存储代价是不能忍受的. 作为统计信息主要组成部分的直方图信息可表示为矩阵,通过对不同数据集统计信息实验,我们认识到矩阵中的大部分格的值为零,也就是说,矩阵是稀疏的.为此我们可以采用压缩存储的方法减少统计信息的存储代价,即只存储值不为零的格. 这种方法的压缩率非常高,使存储代价与格数增长呈线形关系.

6.2 模式指导代价树生成

采用绝对路径方法生成的模式信息,确定了结点的结构嵌套关系,可以用来判断表达式中的结点是否影响其他结点的选择性,从而跳过不必要的直方图运算.

定理 1. 若路径中某中间结点无谓词约束,无嵌套标记,则该结点对其祖先、后代结点的选择性为 100%.

证明略.

定理 2. 若路径表达式某中间结点 n_2 无谓词约束,有嵌套标记,则在形如路径 $\phi n_1[p_1]//n_2/n_3[p_3]$ 中对祖先,或在 $\phi n_1[p_1]/n_2//n_3[p_3]$ 中对后代的选择性为 100%. 证明略.

根据上述定理,SGM 方法在构造 CT 时,通过对结点谓词和嵌套标记的判断,跳过那些对其他结点选择性为 100%的中间结点.

假设路径中结点均无嵌套标记,SGM 方法生成例 2 和例 3 的 CT 如图 6(a),(b)所示. 与图 4(a)比较减少运算 5 次. 采用 SGM 方法生成例 3 的 CT 与图4(b)比较减少运算 3 次.

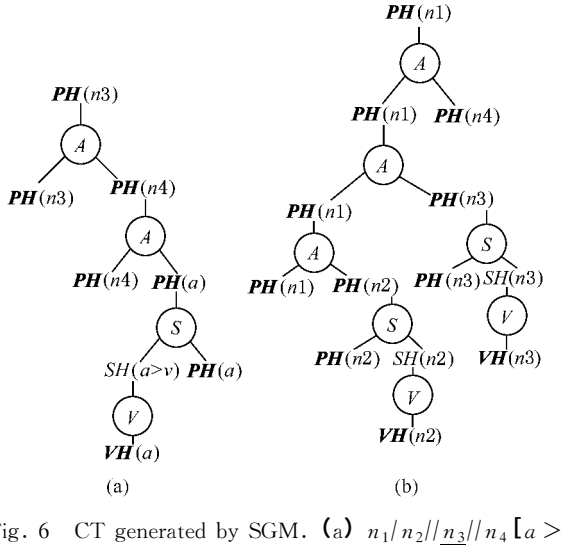

Fig. 6 CT generated by SGM. (a) $n_1/n_2//\underline{n_3}//n_4[a>v]$ and (b) $\underline{n_1}[n_2>v][n_3>w]//n_4$.

图 6 SGM 方法生成 CT.(a) $n_1/n_2//\underline{n_3}//n_4[a>v]$;(b) $\underline{n_1}[n_2>v][n_3>w]//n_4$

7 实验和算法分析

我们在 Win XP 中用 VC++ 编码实现本文提出的方法,实验机器的 CPU 为 PⅣ 1.6GHz,内存为 256MB. 测试集有两个,一个是 Xmark[10],数据结构嵌套关系复杂,数据和结构分布相对合理(相关性

不大). 另一个是由 XML SPY[11] 生成 Bib 数据集,数据模式含 11 个不同类型的结点,其中 6 个结点包含值. 结构相对简单,数据和结构分布扭曲,并且其结点值的分布高度相关.

7.1　统计信息存储代价

Bib 数据集在不同情况下的统计信息存储代价如图 7 所示,存储代价最小的是压缩 PLH + VH 方法. 其次是压缩 PH + VH 方法和无压缩的 PLH + VH 方法,成线性发展. 无压缩的 PH + VH 在方法存储代价稍高. XSketch 中常用的值分布直方图为二维. 对 6 个两两相关的值建立二维直方图的代价随格数平方级增长,当格数为 100×100 时,达到 625KB,与无压缩的 PH + VH 接近. 当用三维直方图统计相关值之间的相关性时,其存储代价远远超过无压缩的 PH + VH 方法. 当格数为 30×30 时,其存储代价超过 2MB.

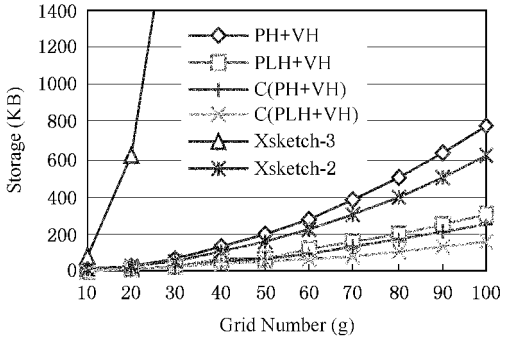

Fig. 7　Storage cost.

图 7　存储代价

7.2　选择性计算代价

我们设计了两组不同特征的查询,一组为单个谓词,但路径较长(>4). 另一组路径较短,但含值谓词较多(>2). 每组查询有 5 个实例. 我们以 2M 规模 XMark 数据集查询时间为基准计算代价估计时间与查询执行时间之间的比值.

图 8 比较了 PM,SGM 与二维直方图 XSketch 方法的对上述两组查询的平均代价计算效率. 与 PM 方法相比,由于有效地减少了结构选择计算,SGM 方法的效率提高一倍. 而 XSketch 方法无需多个直方图运算,计算代价较小.

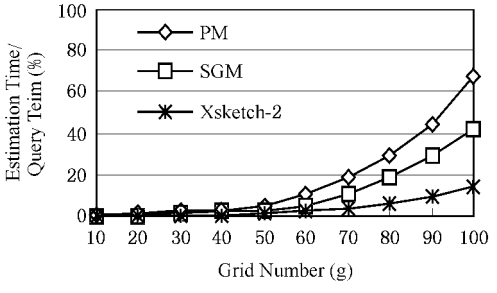

Fig. 8　Selectivity estimation cost.

图 8　选择性计算代价

7.3　选择性计算精度

我们采用相对误差计算方法. 查询数据集采用 2MB 数据集和 10MB XMark 数据集. 图 9 显示路径表达式中只有单个值谓词时的计算误差情况. 由于准确地抓住了结构与值之间的相关性信息,无论在数据分布非常扭曲的 Bib 数据集,还是在数据分布相对规则的 XMark 数据集,PH 方法和 PLH 方法在格数大于 15×15 时均能准确地估计查询选择性. 而这时存储空间不过 30KB,代价估计的时间不到查询时间的 1%,说明了本文方法的高效性和准确性. 而 XSkerch 二维直方图方法的误差很大,这在 Bib 数据集上表现得非常明显. 而且格数的增长不能有效减少误差.

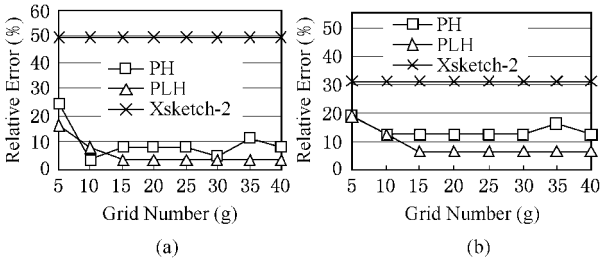

Fig. 9　Estimation accuracy with one predicate. (a) Bib DataSet and (b) XMark DataSet.

图 9　单个谓词表达式选择性计算精度. (a) Bib 数据集;(b) XMark 数据集

当格数很小时(<10),直方图计算误差较大,尤其是 PH 方法,这是由于格内计算误差造成的. 经过大量的实验我们发现,把格数控制在 20 左右能够获得存储、效率和准确性的最佳组合.

多个谓词的代价估计误差如图 10 所示. PH 与

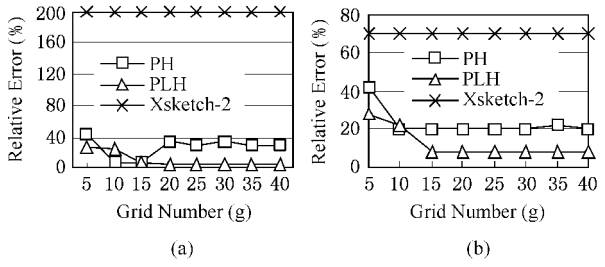

Fig. 10　Estimation accuracy with multi-predicates. (a) Bib DataSet and (b) XMark DataSet.

图 10　多个谓词表达式选择性计算精度. (a) Bib 数据集;(b) XMark 数据集

PLH 方法的平均误差较单谓词路径相比略有增长,其中,PH 方法增长较明显. 而 XSketch 基于独立性分布假设,导致误差的成倍增长.

8 总 结

XML 数据中普遍存在结构和数值的相关性. 以前的研究工作采用独立性分布假设,割裂地讨论结构相关和值相关问题,是计算误差的根本原因. 本文提出一种直方图计算方法,实验证明本文提出的方法在数据分布扭曲和数据结构复杂的情况下,均能获得有效精确的估计结果. 这种方法既可用于最终结果大小的估计,也可用于计算中间结果选择最优执行计划.

参 考 文 献

1 J. McHugh, J. Widom. Query optimization for XML. In: Proc. 25th VLDB Conf. San Francisco: Morgan Kaufmann, 1999. 315~326

2 A. Aboulnaga, R. A. Alameldeen, J. Naughton. Estimating the selectivity of XML path expressions for internet scale applications. In: Proc. 27th VLDB Conf. San Francisco: Morgan Kaufmann, 2001. 591~600

3 Z. Chen, V. H. Jagadish, F. Korn, *et al*. Counting twig matches in a tree. In: Proc. 17th ICDE Conf. Los Alamitos, CA: IEEE Computer Society Press, 2001. 595~604

4 N. Polyzotis, M. Garofalakis. Statistical synopses for graph-structured XML databases. In: Proc. 2002 ACM SIGMOD Conf. New York: ACM Press, 358~369

5 J. Freire, R. Jayant, M. Ramanath, *et al*. StatiX: Making XML count. In: Proc. 2002 ACM SIGMOD Conf. New York: ACM Press, 2002. 181~191

6 Y. Wu, J. Patel, H. Jagadish. Estimating answer sizes for XML queries. In: Proc. 8th EDBT Conf. Berlin: Springer, 2002. 590~608

7 W. Wang, H. Jiang, H. Lu, *et al*. Containment join size estimation: Models and methods. In: Proc. 2003 ACM SIGMOD Conf. New York: ACM Press, 2003. 145~156

8 N. Polyzotis, M. Garofalakis. Structure and value synoposes for XML data graphs. In: Proc. 28th VLDB Conf. San Francisco, CA: Morgan Kaufmann, 2002

9 N. Polyzotis, M. Garofalakis, Y. Ioannidis. Selectivity estimation for XML twigs. In: Proc. 20th ICDE Conf. Los Alamitos, CA: IEEE Computer Society Press, 2004

10 CWI. Xmark. http://monetdb.cwi.nl/xml, 2003-01

11 ALTOVA, Inc. XML Spy. http://www.xml.com/pub/p/15, 2004

Wang Yu, born in 1973. Ph. D. Her main research interests include Web data management and XML database.

王宇,1973 年生,博士,主要研究方向为 Web 数据管理、XML 数据库.

Meng Xiao Feng, born in 1964. Professor and Ph. D. supervisor. His main research interests include Web data integration, XML database and mobile database.

孟小峰,1964 年生,教授,博士生导师,主要研究方向为 Web 数据集成、XML 数据库、移动数据管理等.

Wang Shan, born in 1944. Professor and Ph. D. supervisor. Her main research interests include data warehousing and business intelligent technology, mobile database system and Web data management.

王珊,1944 年生,教授,博士生导师,主要研究方向为数据仓库、商务智能、移动数据库、Web 数据管理等.

Research Background

This work is supported by the 863 High Technology Foundation of China under grant number 2002AA116030, the Natural Science Foundation of China (NSFC) under grant number 60073014, 60273018, the Key Project of Chinese Ministry of Education (No.03044), and the Doctoral Foundation of Hebei University.

As XML becomes the standard for data exchanging over the Internet, much research has been devoted to the efficient support of XML queries. Compared with traditional query optimization technologies, XML query optimization has complex data model, weak schema information supporting, and insufficient relative basic research. So it needs some particular technologies. An XML query, expressed by a path expression, may contain branches with predicates, and each branch may have different impact on the selectivity of the entire query. If the branch with less selectivity is queried first, the intermediate result will have a relatively small size and the query efficiency can be improved. XML data is a hybrid of structures and values, some of which are highly correlated. Previous methods for selectivity estimation deal with the distribution of the value and the structure by using independent assumption, and have not taken into consideration the predicate value correlation among the different branches. We design a novel method connecting the isolated estimation together like a bridge. The method is of great accuracy especially when the distribution of the value or structure of the data is very skew.