

XML 数据扩展前序编码的更新方法*

罗道锋⁺, 孟小峰, 蒋瑜

(中国人民大学 信息学院, 北京 100872)

Updating of Extended Preorder Numbering Scheme on XML

LUO Dao-Feng⁺, MENG Xiao-Feng, JIANG Yu

(Information School, Renmin University, Beijing 100872, China)

+ Corresponding author: Phn: +86-10-62515575, E-mail: jiangyu@ruc.edu.cn, <http://www.ruc.edu.cn>

Received 2003-12-22; Accepted 2004-03-31

Luo DF, Meng XF, Jiang Y. Updating of extended preorder numbering scheme on XML. *Journal of Software*, 2005,16(5):810-818. DOI: 10.1360/jos160810

Abstract: Most of the XML query processing strategies are based on some numbering schemes. Nodes on the XML tree will be assigned a unique code by the numbering scheme, and ancestor-descendant relationship could be directly told through the codes. The most famous numbering scheme is Region Based Numbering Scheme. However, XML data will be updated. Once the data is updated, the region code should be adjusted to keep the indexing and query processing techniques working. Unfortunately, few studies have been reported on the issue of the numbering scheme. This paper focuses on this issue, proposing a series of space preserving and updating algorithms. Extensive experiments are conducted to test the effectiveness of the algorithms.

Key words: XML; numbering scheme; region code; update; reserve

摘要: 大部分 XML 查询技术都是基于某种对 XML 树的编码方法。对 XML 树的编码,是指按照某种规则对 XML 树的每一个结点分配唯一的编码,目的是通过任意两个结点的编码,能够直接判断两个结点之间是否具有祖先后代关系。最常用的编码方法是区域编码方法(region based numbering scheme)。然而,XML 数据也会面临插入删除等更新问题。数据一旦更新,区域编码也要作相应的调整,才能保证基于这个编码的各种索引和查询算法的正确性。在编码的更新方面,目前研究得还不多。主要研究区域编码的更新问题,采用预留编码空间的方法,针对不同特征的 XML 数据和应用环境提出了一整套预留算法和编码更新算法,并做了大量的实验,检验这些算法的有效性。

关键词: 可扩展标记语言;编码方案;区域编码;更新;预留

中图法分类号: TP311 文献标识码: A

* Supported by the National Natural Science Foundation of China under Grant No.60273018 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2002AA116030 (国家高技术研究发展计划(863)); the Key Project of Chinese Ministry of Education under Grant No.03044 (国家教育部科学技术研究重点项目)

作者简介: 罗道锋(1978—),男,广东河源人,硕士,主要研究领域为 XML 存储和查询;孟小峰(1964—),男,博士,教授,博士生导师,主要研究领域为 XML 数据库,Web 数据集成和移动数据库技术;蒋瑜(1981—),男,硕士,主要研究领域为 XML 查询,索引技术。

随着XML的迅速发展,对XML数据的高效索引和查询的需求越来越迫切.近年来,人们提出了各种各样的关于XML数据的索引和查询技术,如EE-Join,EA-Join和KC-Join^[1],MPMCJN^[2],tree-merge,stack-merge^[3],XPath Accelerator^[4],Containment Join size Estimation^[5],等等.这些技术大部分基于某种对XML树的编码方法^[1,2,6-9],其中最为流行的一种是基于区域的编码方法(region based numbering scheme)^[1,2,6,7].

在基于区域的编码方法中,树中的每个结点都被赋予一对数字,这对数字表达这个结点所覆盖的区域.如果一个结点的区域包含另一个结点的区域,则在树中前者是后者的祖先.在文献[2,6]中,每个结点以它在树的前序遍历顺序中的开始(start)和结束(end)编号作为编码.这种编码方法最大的缺点是缺乏灵活性,一旦插入新的子树,整个树不得不重新编码.为了解决这个问题,文献[1]提出了一种扩展的前序编码方法(extended preorder numbering scheme),每个结点都用一个 $(order, size)$ 对来标识,其中,order是该结点在树中的前序遍历顺序,size是以该结点为根的子树的结点数.对于给定结点 T_1 和 T_2 , T_1 是 T_2 的祖先,当且仅当 $T_1.order < T_2.order$ 且 $T_1.order + T_1.size > T_2.order + T_2.size$.该编码方法在编码时给结点的size值比它实际的大小要大,这样,就可以容纳将来新插入的结点,而不影响其他结点的编码.

当XML文档发生更新时(如插入1棵子树),需要相应地调整结点的编码以维持其反映祖先-后代关系的特性.这种重新编码的代价可能是很大的,有时甚至需要更新整棵树的编码.减少编码更新代价的关键在于怎样预留编码空间和更新发生时如何重新编码.但是,就我们所知,在这个问题上研究得还比较少.尽管文献[1]曾提到通过给结点一个比实际大小更大的一个size来容纳将来的插入,但是它并没有涉及如何预留以及将来如何使用预留的空间的问题.本文将就区域编码的预留和更新问题进行讨论.文献[1]的主要贡献如下:

- 提出了一整套编码预留算法.不同的数据和应用,对编码预留有不同的影响.本文提出了在模式独立、基于数据模式和基于更新模式3种情况下的编码预留算法.

- 提出了编码的单次更新算法,并在此基础上进一步提出了批量更新算法.

- 我们设计并实现了在各种不同情况下的编码更新,并且通过实验展示了以上几种方法的有效性.

本文第1节详细分析影响编码预留和更新的因素.第2节针对影响编码预留的因素讨论不同的编码空间预留算法.第3节阐述编码的单次更新和批量更新算法.第4节是实验结果.第5节总结全文和展望下一步工作.

1 影响编码预留的因素

在1棵新的子树(inserted subtree)插入到1棵XML目标树(target tree)中之前,我们需要确定子树的插入位置(insert position).在这里,我们用元组 $(parent, childIndex)$ 表示插入位置,其中parent表示inserted subtree的父亲,childIndex表示它在其父亲的孩子中的编号(第1个孩子编号是0).

在编码的预留和更新中,有下面几个考虑因素.

1.1 数据模式

在XML的更新处理中,数据模式扮演了很重要的角色.根据数据有没有模式约束,数据可以分为两种:基于模式的数据和模式独立的数据.对基于模式的数据,模式信息(如DTD或者XML Schema)对插入的子树有所限制,因为插入子树之后的数据,仍需要符合模式定义;而对模式独立的数据来说,XML数据的插入没有任何限制,它可以发生在树中的任何可能的插入点.

1.2 更新模式

在实际中,各种结点插入的频率是不相同的.根据结点插入的频率,可以为每一个结点赋一个插入权重.权重越高,表示插入的频率越高,插入的可能性越大.我们把结点的插入权重叫做更新模式(update pattern).就编码更新这个问题而言,更新模式比数据模式更能有效地预留编码空间,因为数据模式只是告诉我们可以或者不可以,而更新模式更具体地告诉我们大概插入多少.因此,如果有更新模式支持,预留算法将会更行之有效,更新代价会更小.

1.3 批量更新

在数据仓库的应用中,数据是在某个时间段批量插入的.类似地,在 XML 数据中,这种情况同样存在.这时,对编码的更新是批量的.关键问题是给新插入的子树分配编码的顺序.一个可能的方法是按照数据插入的顺序,先插入的子树先分配编码;或者按照插入的数据的前序遍历顺序,前序遍历靠前的新插入子树先分配编码.但是,我们将会看到,这些方法最大的问题是它会带来重复编码,即某棵子树已经新分配了编码,后来由于其他子树的编码,需要再一次更新它的编码.在批量更新的情况下,如何避免重复编码,减少更新代价,也是一个重要的问题.

本文将就这 3 个问题逐步深入地展开讨论.为了简单起见,我们使用 DTD 作为模式信息.不失一般性,本文简单地认为 XML 数据是一棵由结点组成的树.考虑到删除并不影响编码的更新,我们只集中讨论数据插入的情况,不考虑插入可选结点的情况.

2 编码空间的预留

在扩展前序编码方法中,通过将节点 $(order, size)$ 对中的 $size$ 设得比实际的 $size$ 大来预留编码空间.为了以示区别,称编码中的 $size$ 值为 $regionSize$,而结点的实际 $size$ 值为 $actualSize$.在这一节里,从简单到复杂,我们分别讨论模式独立的数据、有数据模式的数据和有更新模式的数据的编码预留方法.

2.1 模式独立数据的预留算法

模式独立的数据,对新插入子树的类型和插入的位置没有任何限制.在这种情况下,我们采取“平均预留”策略:把编码空间平均预留到所有可能的插入位置.为了最大限度地利用编码空间,显然,目标树的根结点的 $regionSize$ 应该被赋值为整个编码空间的大小.比如,如果 $regionSize$ 的长度是 4 个字节,那么根结点的 $regionSize$ 应该是 2^{32} .

平均预留策略是把编码空间平均预留到所有可能的插入位置.

定理 1. 一棵模式独立的 XML 树,假设它的 $size$ 是 s ,则插入位置的个数是 $2s-1$.

根据平均预留策略,在图 1(a)中,存在 15 个插入位置,假设最大编码空间大小 $maxSize=100$,我们将为每个插入位置预留 $(100-8)/15=6.13$ 的空间.模式独立数据的预留算法很简单,限于篇幅,本文略去这个具体算法.在实际计算 $order$ 和 $size$ 的时候,我们采用浮点数,只是在最后赋值的时候取整,这样就能有效地避免计算中的累计误差.

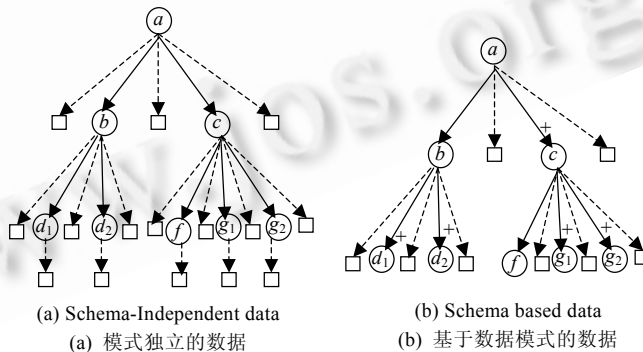


Fig.1 Data pattern
图 1 数据模式

2.2 基于数据模式的预留算法

基于数据模式的 XML 数据,对新插入子树的类型和插入位置是有限制的.我们可以预测新插入子树与其兄弟子树在结构上是相似的.所以,我们的策略是为一个插入位置预留的空间大小,等于该插入位置两旁的兄弟结点大小的一个倍数.为了描述我们的算法,首先引入一个概念——预留因子(reserving factor).

定义 1. 一个 size 为 s 的结点,如果它的预留因子是 α ,则它对父亲的 regionSize 的贡献是 $s \times \sigma$.

假设一个结点的预留因子是 σ ,它对父结点的贡献是 $\sigma \times \text{regionSize}$,其中, $(\sigma-1) \times \text{regionSize}/2$ 作为预留空间,分别预留在了该结点的左右两边的位置.以图 1(b)中的树为例,我们假设所有的可重复结点 c, d_1, d_2, g_1 和 g_2 都具有相同的预留因子 σ ,则 d_1 对 b 的 regionSize 的贡献是 σ, b 的 regionSize 是 $1+2 \times \sigma$.类似地, c 的 regionSize 是 $2+2 \times \sigma, a$ 的 regionSize 是 $1+(1+2 \times \sigma)+(2+2 \times \sigma) \times \sigma=2+4\alpha+2\alpha^2$.现在我们给定最大编码空间 $\text{maxSize}=100$,则可以得到方程:

$$2+4\alpha+2\alpha^2=100 \tag{1}$$

解得 $\sigma=6.07$.求解 σ 的方程具有如下的一般形式:

$$s_0+s_1\alpha+s_1\alpha^2+\dots+s_n\alpha^n=\text{maxSize}.$$

求解预留因子 σ 的关键是得到这个方程的系数 s_i .

定义 2(可重复结点的层次). 一个可重复结点的层次是从树根到这个结点的路径上的所有可重复结点的个数.

如在图 1(b)中 c 的可重复层次是 1,而 g_1 的是 2, a 的是 0.

定义 3(可重复结点的本层次大小(pure size)). 假设 e 是一个可重复结点, size 为 s ,而 d_1, d_2, \dots, d_n 是 e 的子孙可重复结点,且它们都处在 e 的下一个可重复层次,它们的 size 分别是 s_1, s_2, \dots, s_n ,则 e 的本层次大小是 $ps(e)=s-(s_1+s_2+\dots+s_n)$.

c 的 size 是 4, c 的后继中 g_1 和 g_2 都是它下一级的可重复结点,且 size 都是 1,则 $ps(c)=4-1-1=2$.在这棵树中,可重复级别是 1 的结点的本层次大小之和等于 2,恰好等于式(1)中一次项的系数.不难看出:

定理 2. 假设 e_1, e_2, \dots, e_n 是目标树中所有可重复层次为 i 的结点.在求解预留因子的方程中,第 i 次项的系数 S_i 等于 $\sum_{j=1}^n ps(e_j)$.

这个定理的证明此处从略.在求出式(1)的各项系数后,我们采用二分法求解 α .计算出预留因子后,给结点进行编码过程就比较简单了.如图 2 所示.

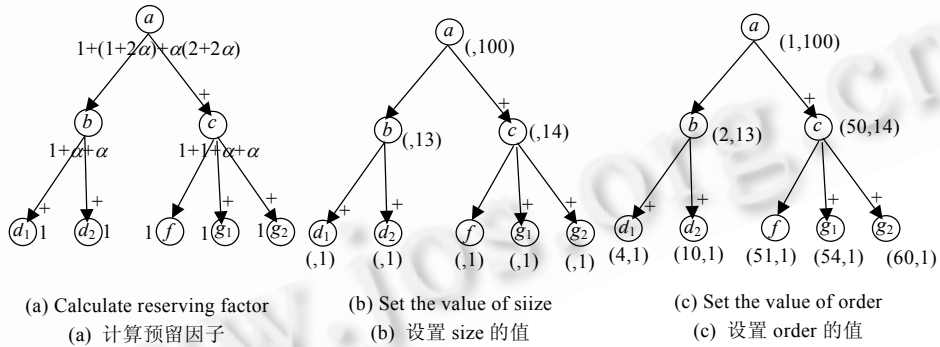


Fig.2 Schema based space reserving algorithm
图 2 基于模式数据的空间预留算法

2.3 基于更新模式的预留算法

数据模式只是告诉我们某一个插入位置可以或者不可以插入 1 棵子树,却不能告诉我们插入多少.因此,每一个可重复结点的预留因子都是 σ .更新模式却能准确地反映每一个插入位置可能插入新的子树的概率.根据更新模式,每一个可重复结点的预留因子就有了一个权重.

更新模式可以由统计数字获得.假设在一段时期内,结点类型 $E_1, E_2, E_3, \dots, E_n$ 的结点分别被插入了 $p_1, p_2, p_3, \dots, p_n$ 次($p_1 \leq p_2 \leq p_3 \leq \dots \leq p_n$).结点类型 $E_1, E_2, E_3, \dots, E_n$ 将被赋一个插入权重 $w, w_1=1, w_2=p_2/p_1, w_3=p_3/p_1, \dots, w_n=p_n/p_1$.

假设预留因子是 σ ,对于一个插入权重为 w 的结点来说,它对父结点的贡献是 $w \sigma \times \text{regionSize}$,其中, $(w\sigma-1) \times \text{regionSize}/2$ 作为预留空间,分别预留在了该结点的左右两边的位置.

同样地,为了求解预留因子 σ ,需要构造方程:

$$s_0 + s_1 \alpha + s_1 \alpha^2 + \dots + s_n \alpha^n = \text{maxSize}.$$

与第 2.2 节的分析类似,给出如下定理.

定理 3. 假设 e_1, e_2, \dots, e_n 是目标树中所有可重复层次为 i 的结点. 对于 $e_j (1 \leq j \leq n)$, 从根结点到 e_j 的路径上的可重复结点的插入权重分别为 w_1, w_2, \dots, w_i . 令 $W_j = \prod_{k=1}^i w_k$. 在求解预留因子的方程中, 第 i 次项的系数 S_i 等于

$$\sum_{j=1}^n ps(e_j) W_j.$$

在有更新模式下的求解预留因子算法, 设置 `regionSize` 算法和设置 `regionOrder` 算法均与第 2.2 节的算法类似, 限于篇幅, 这里不再详述.

3 编码更新策略

在第 2 节中, 我们针对模式独立、基于数据模式和基于更新模式的 3 种不同类型的数据讨论了不同的编码空间的预留方法. 当预留空间不足以容纳新插入的子树时, 就需要调整已有结点的编码, 以腾出空间来容纳新的子树. 重新编码的目标是能容纳新插入的子树, 而且更新代价最小. 所谓更新代价, 是指在新子树插入后为它编码的过程中需要重新编码的结点的个数, 不包括新的子树, 也不包括在新子树编码过程中被访问但是没有改变编码的节点个数. XML 数据的编码的存储方式可能是多样化的, 比如和数据存储在一起或者按照某种顺序单独存储, 这不属于本文讨论的范围, 这里从略. 因此, 我们这里仅集中关注上面定义的逻辑层次上的更新代价. 在这一节里, 我们从简单到复杂, 分别讨论单次插入和批量插入情况下的编码更新策略.

3.1 单次插入的编码更新策略

回忆前面提到的插入位置的定义 $(\text{parent}, \text{childIndex})$. 我们假设在插入位置 (p, t) 刚插入了 1 棵子树; p 是新子树树根的父亲结点, 它的所有孩子结点依次为 $c(0), c(1), \dots, c(t), \dots, c(n), c(t)$ 就是刚插入子树, 为了描述方便, 我们还假设 p 有两个虚拟的孩子结点 $c(-1)$ 和 $c(n+1)$, 令

$$\begin{aligned} c(-1).\text{regionOrder} &= p.\text{regionOrder}; \\ c(n+1).\text{regionOrder} &= p.\text{regionOrder} + p.\text{regionSize}; \\ c(-1).\text{regionSize} &= c(n+1).\text{regionSize} = 0; \\ c(-1).\text{actualSize} &= c(n+1).\text{actualSize} = 0. \end{aligned}$$

在 $c(t)$ 子树插入完成后, 我们需要为 $c(t)$ 子树进行编码, 这时最简单的情况就是 $c(t)$ 处的预留空间大小超过 $c(t).\text{actualSize}$, 我们不必对其他结点重新编码就可以完成新子树的编码工作 (判断条件是 $c(t+1).\text{regionOrder} - c(t-1).\text{regionOrder} - c(t-1).\text{regionSize} \geq c(t).\text{actualSize}$). 但是, 如果预留空间不够用, 我们就需要对部分结点重新编码. 这件事情可以分成两步:

步骤 1. 找出需要重新编码的结点集合.

这时又可以分为两种情况:

情况 1. $p.\text{regionSize} \geq p.\text{actualSize}$.

请注意, 这时 $p.\text{actualSize}$ 包含 $c(t).\text{actualSize}$. 显然, 在这种情况下, p 结点不需要重新编码就可以完成这次插入的编码更新. 这时存在着这样的 p 的秩序的孩子列表 $c(i), \dots, c(t), \dots, c(j)$, 其中 $-1 < i \leq t$ 且 $t \leq j < n+1$. 这个列表满足:

$$c(j+1).\text{regionOrder} - c(i-1).\text{regionOrder} - c(i-1).\text{regionSize} \geq \sum_{k=i}^j c(k).\text{actualSize}.$$

因此, 我们只要对整个列表中的结点的子树进行编码即可, 它的更新代价为

$$\text{Renumbering Cost}(i, j) = \sum_{k=i}^j c(k).\text{actualSize} - c(t).\text{actualSize}.$$

对于 $c(t)$, 可能存在许多个这样的列表, 为了最小化更新代价, 我们选取其中代价最小的一个列表作为重新编码的列表.

情况 2. $p.regionSize < p.actualsize$.

对于这种情况,我们必须在 p 结点子树以外的地方寻找更多的预留空间.我们采用了一种巧妙而且有效的方法:将 p 作为新插入的子树,然后试图给它进行编码,这就回到了最开始的情况下,再重新判断.

经过步骤 1 的处理,我们得到一个顺序的孩子队列,接下来是为这个队列重新编码.

步骤 2.为更新队列进行编码.

确定编码空间之后,只剩下如何为更新队列编码的问题.考虑到可能的插入,在对队列完成编码的同时需要在它的内部和两旁作编码空间预留.最好的方法是能够调用第 2 节中的预留算法,给更新队列重新预留空间.为此,我们引入一个特殊的虚结点 `virtualParent` 作为更新队列中的结点的父结点,然后对这棵树编码.最后,删除虚结点 `virtualParent`,将更新队列还原到 p 下面的正确位置,并且保持它们的编码不变.这就完成了对更新队列的重新编码($c(t)$ 包含在队列中,所以它的子树的编码也随之完成).

3.2 批量插入的编码更新策略

批量插入的编码更新策略,可以看做是单次插入的累加.一个简单的方法是按照插入的顺序,给新插入的子树编码;或者按照前序遍历的顺序给新插入子树编码.但是,不恰当的顺序将会导致重新编码,因而增加了编码更新代价.导致重新编码的原因有 3 个:

- (1) 某棵子树被分配了编码,后来由于祖先结点的原因,被重新编码.
- (2) 某棵子树被分配了编码,后来由于兄弟结点的原因,被重新编码.
- (3) 当前父结点已经不足以容纳新的子树,根据第 3.1 节的策略,父结点被当作新插入的子树处理,导致父结点的兄弟结点被重新编码.

在处理批量插入的时候,为了避免重复编码,我们需要 3 个步骤:预处理,恰当的顺序和改进的搜索更新队列的算法.

步骤 1.检查新插入子树的父结点能否容纳新插入的子树,如果不能,把父结点当作新插入结点.这样做的目的是先把情况 3 去除,保证每个新插入子树的父结点都能容纳新插入的子树.

步骤 2.按照广度优先的顺序对批量插入的子树进行排序.这样可以避免情况 1 的发生,因为按照广度优先的顺序,祖先结点总是先于子孙结点被处理.

步骤 3.改进的搜索更新队列算法用来避免情况 2.在单个新插入子树的处理中,我们只为该新插入的子树寻找更新队列.为了避免情况 2,我们需要为同一个父结点下的所有新插入子树一起寻找更新队列.寻找的结果是若干条不交叉的更新队列,它们包含了该父结点下所有新插入子树.

经过这 3 个步骤,可以完全避免重复编码.

4 实验

考虑到不同的 XML 数据有特定的 tree 结构,它们在实验中的表现也会不一样,我们在实验时选取了两个特征突出的数据集 Xmark^[10]和 Shakespeare^[11].

4.1 实验的基本过程和系统参数

首先我们列出可能影响更新代价的系统参数,见表 1.

Table 1 System parameters of experiment

表 1 实验的系统参数

Parameter name	Annotation	Default value
ST	Initial size of the target tree	33 140
UR	Maximal update ratio	220%
LEN	Code length	64

ST 是初始的目标树的大小(initial size of the target tree),LEN 是编码的长度,它是(order,size)对中 order 和 size 的长度之和,它也直接确定了编码空间的大小.如 LEN=64,则 order 和 size 的取值范围都是 $(0 \sim 2^{32-1})$,即编码空间大小为 2^{32} .UR 是最大更新比,是指在所有插入完成后目标树的大小与初始大小 ST 的比.在以后的各个图表

中,系统参数如果没有特别说明,都取默认值.在各个实验结果的图示中,更新代价(update cost)采用 UC 的简写形式.

下面是在实验过程中一次数据插入过程的描述,它由以下的 3 个步骤组成:

首先是新插入子树(inserted subtree)的生成.我们用一个称为源树的 XML 文档,用来生成新插入的子树.它与目标树是由同一个 XML 数据集生成的,所以它们具有相同的模式定义.接着是插入位置的选取.插入位置由对(parent,childIndex)表示.最后一步,把新插入子树插入到插入位置.这在第 3 节已有详细讨论,这里不再多加说明.实验时,我们不断地重复以上的 3 个步骤,直到某一次插入完成后目标树的大小达到预定的值 $ST \times UR$,然后计算更新代价.对于每次的实验,我们都重复 10 次取平均作为最终的实验结果.

4.2 模式独立的数据的编码更新代价

由于没有模式信息的支持,我们在插入时只能随机选择插入子树和插入位置.考虑在树根的附近插入一棵庞大的子树,这时可能需要对整棵树重新编码,在多次插入后总的更新代价必然是难以想象的.由于新插入子树和插入位置的完全随机性,更新代价也呈现很大的随机性和不稳定性.图 3 反映了同一次实验反复进行了 10 次的更新代价的情况,结果表现出极端的不稳定性.由于这种不稳定性,我们下面的实验集中在基于数据模式和更新模式的数据上.

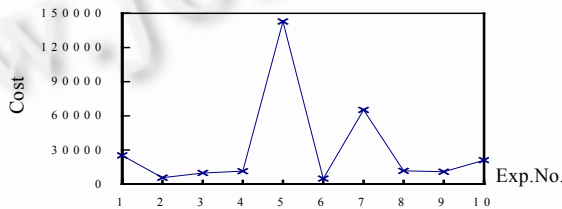


Fig.3 UC of schema-independent data (ST=17132)
图 3 模式独立数据的更新代价(ST=17132)

4.3 基于数据模式的数据的编码更新代价

在基于数据模式的数据的编码中,LEN 和 UR 对更新代价的影响是很明显的.更大的 LEN 意味着更大的编码空间,从而会降低重新编码的结点数;而 UR 的增大必然导致更新代价的增大.图 4 说明了 LEN 和 UR 更新代价的影响.

再看图 5,图中的横坐标是 ST,纵坐标是更新比=更新代价(Cost)/ST.在实验时我们用不同的文档在不同的 LEN 下进行了多次实验,图 5 中的数据只是其中的一组.可以看出,图中更新比的值基本上固定在 18%左右,可以认为 ST 的大小对更新比的影响很小.

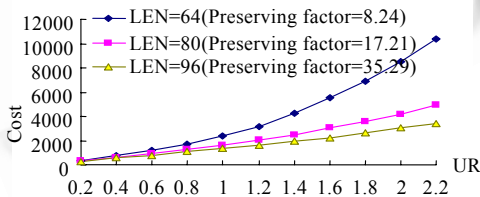


Fig.4 UC under different LEN (Cost)
图 4 不同 LEN 下的更新代价(Cost)

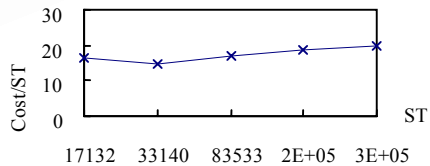


Fig5. UC ratio under different ST (LEN=80)
图 5 不同 ST 下的更新代价比(LEN=80)

为了说明我们的预留和更新算法的效果,我们做了表 2 的统计,表中的粗体字是指对应情况发生的次数和占总次数的百分比.在实验中,当 UR 达到 220%时,大概完成了 5 280 次插入操作,其中 80%以上的插入都没有进行重新编码,这从一定程度上说明我们的算法是高效的.

Table 2 Analysis of UC (ST=33140,UR=220%)

表 2 更新代价分析(ST=33140,UR=220%)

		LEN=64	LEN=80	LEN=96
Case 1	Cost=0	4332 (82%)	4629 (87%)	4706 (90%)
	Cost>0	417 (8%)	195 (4%)	128 (2%)
	Cost	6 753	2 632	1 514
Case 2	Percentum	531 (10%)	473 (9%)	431 (8%)
	Cost	3 953	2 303	1 891
Sum of update cost		10 346	4 935	3 405

4.4 文档特征对更新代价的影响

在基于模式的数据前提下,扁平的树结构比相对较深的树结构有更好、更多的表现.图 6 的实验数据集是 Shakespeare, 树结构比较扁平,ST=28300;而图 4 的数据集是 Xmark,树结构较深且包含递归定义的结点,ST=33140.它们初始的 ST 相差不大,但是两个图中的曲线(均对应 LEN=64 的情况)指的更新代价相差有 100 倍.这是因为在图 3 中,LEN=64 时的预留因子是 8.24,而在图 6 中,同样 LEN=64,但是预留因子达到 20.59.

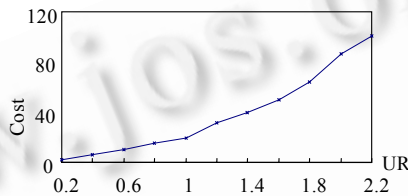


Fig.6 Impact of data character (ST=28300, LEN=64)

图 6 文档特征的影响(ST=28300,LEN=64)

4.5 基于更新模式的数据的编码更新代价

在具有更新模式的情况下,预留空间将只分配给具有更新权重的节点的子树,而且根据各自的更新权重的不同,不同的可重复节点得到的预留空间也会不同.在 Xmark 数据集中,考虑到实际应用中绝大部分的插入都集中在 item 等几种节点的子树的需求上,我们设计了有更新模式与没有更新模式的更新代价对比的实验,图 7 是实验的结果,在实验中我们的更新模式设定是:open_auction:30,closed_auction:30,person:40,item:50,bidder:50.对于其他的可重复节点,我们设定更新权重均为 0.可以看出,相对于没有更新模式的情况,更新代价有了大幅度的削减.

4.6 批量更新的代价

大量的单个子树更新带来了大量的重复编码,使编码更新代价很大.采用批量更新,避免了重复编码,相比之下,编码代价小了很多.图 8 表示各种更新率下单次批量更新和多次单棵子树更新的代价比较,其中单棵子树的更新采用的是基于数据模式的预留方法.比如,在 UR=1.0 的情况下,代价是 504,表示一次批量插入 1.0×33140 个结点,然后批量更新,编码更新的结点数是 504.可以看出,批量更新的代价相比同更新率的多次单棵子树更新的代价,要小得多.

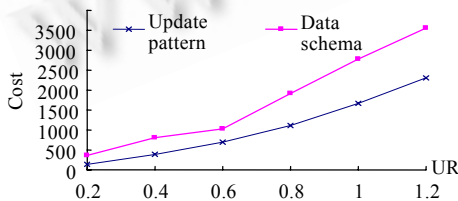


Fig.7 Update pattern v.s. data schema

图 7 基于更新模式和基于数据模式的代价比较

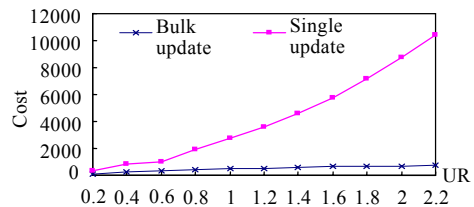


Fig.8 Bulk update v.s. multi single update

图 8 BulkUpdate 和多次 single update 的代价比较

5 结论和展望

本文集中讨论了关于 XML 文档的扩展前序编码的更新问题,根据不同的数据和应用,提出了一整套预留和更新算法.实验的结果表明:在各种不同的情况下,这几种算法都表现出较好的性能.这是在 XML 文档编码更新问题上进行的有效尝试.

在本文的基础上,在 XML 数据的编码更新领域还有很多进一步的工作可以展开.比如,由于 XML 数据的编码是为了查询处理的需要,在编码要更新时,应该怎样更新才能够不影响或者尽可能小地影响查询处理的操作?这些都是值得思考的问题.

References:

- [1] Dietz PF. Maintaining order in a linked list. In: Proc. of the 14th Annual ACM Symp. on Theory of Computing. San Francisco, 1982. 122–127.
- [2] Lee YK, Yoo SJ, Yoon K. Index structures for structured documents. In: ACM 1st Int'l Conf. on Digital Libraries. Bethesda, 1996. 91–99.
- [3] Li Q, Moon B. Indexing and querying XML data for regular path expressions. In: Apers PMG, Atzeni P, Ceri S, Paraboschi S, Ramamohanarao K, Snodgrass RT, eds. Proc. of the 27th VLDB. Roma: Morgan Kaufmann Publishers, 2001. 361–370.
- [4] Al-Khalifa S, Jagadish HV, Koudas N, Patel JM, Srivastava D, Wu Y. Structural joins: A primitive for efficient XML query pattern matching. In: Proc. of the 18th ICDE. San Jose: IEEE Computer Society, 2002.
- [5] Wang W, Jiang H, Lu H, Yu JX. PBiTree coding and efficient processing of containment join. In: Proc. of the 19th ICDE. Bangalore, 2003. 391–402.
- [6] Zhang C, Naughton JF, DeWitt DJ, Luo Q, Lohman GM. On supporting containment queries in relational database management systems. In: Proc. of the 27th ACM SIGMOD. Santa Barbara, 2001. 425–436. <http://www.acm.org/sigs/sigmod/sigmod01/e-proceedings/papers/Research-Zhang-et-al.pdf>
- [7] Grust T. Accelerating XPath location steps. In: Franklin MJ, Moon B, Ailamaki A, eds. Proc. of the 28th ACM SIGMOD. Madison, 2002. 109–120.
- [8] Wang W, Jiang H, Lu H, Yu JX. Containment join size estimation: Models and methods. In: Halevy AY, Ives ZG, Doan AH, eds. Proc. of the 29th ACM SIGMOD. San Diego, 2003. 145–156.
- [9] Schmidt AR, Waas F, Kersten ML, Carey MJ, Manolescu I, Busse R. XMark: A Benchmark for XML data management. In: Dayal U, Ramamritham K, Vijayarman TM, eds. Proc. of the 28th VLDB. Hong Kong, 2002. 974–985. <http://www.vldb.org/conf/2002/S30P01.pdf>
- [10] Kha DD, Yoshikawa M, Uemura S. An XML indexing structure with relative region coordinate. In: Proc. of the 17th ICDE. Heidelberg: IEEE Computer Society, 2001. 313–320.
- [11] Shakespeare dataset. <http://www.cs.kuleuven.ac.be/~ml/ie/>