A４－８

# 階層構造を識別可能な木節点の番号付け

佐藤 隆士, 里本 智彦, 小畑 喜平, 潘 洪涛

大阪教育大学大学院総合基礎科学専攻数理情報コース

キーワード：XML-DB, 階層構造, 木節点の番号付け,半構造データ

# 概　要

- はじめに
- 正規経路式(regular path expression)
- 階層構造を識別可能な番号付け
  - 数字の組による節点の位置表現
  - コード番号による節点の位置表現
  - オーバフローの表現
- 位置関係の計算
  - 絶対位置の復元
  - 節点間の関係の計算
- まとめ

# はじめに

- ＸＭＬは，インターネットの標準的なデータ交換手段として用いられるようになってきている．
- ＸＭＬ文書は木で表現できる階層構造をなしている．
- 階層をたどる経路式質問と呼ばれる特徴的な質問がある．
- 経路式質問を効率よく処理するため，経路に基付く索引が提案されている．
- 節点の木における絶対的な位置を記憶し，且つ様々な節点間の関係を容易に計算できる方法を提案する．

# 正規経路式(regular path expression)

- 経路の途中は任意で、階層の上位と下位を指定するなどの質問式
- 木構造の節点に対応する要素，属性を表す節点に番号を振り，索引には，節点ごとに分解して，番号とともに格納する．
- 検索時には，付けられた節点番号から節点間の先祖子孫関係などの情報を得る．
- Liらの方法では，節点対が，先祖子孫関係であるかどうかが分かるが，その関係が親子であるか，あるいは何世代離れた関係であるかに答えることはできない．
- Leeらの方法は，節点に付けられた番号から，木における絶対的な位置がわかるが，番号を木の幅方向に振っているので親子の関係以外の計算は容易ではない．また，多くの利用されない仮想節点に番号を使うため，arityと木の高さが大きくなると，現実的でない．

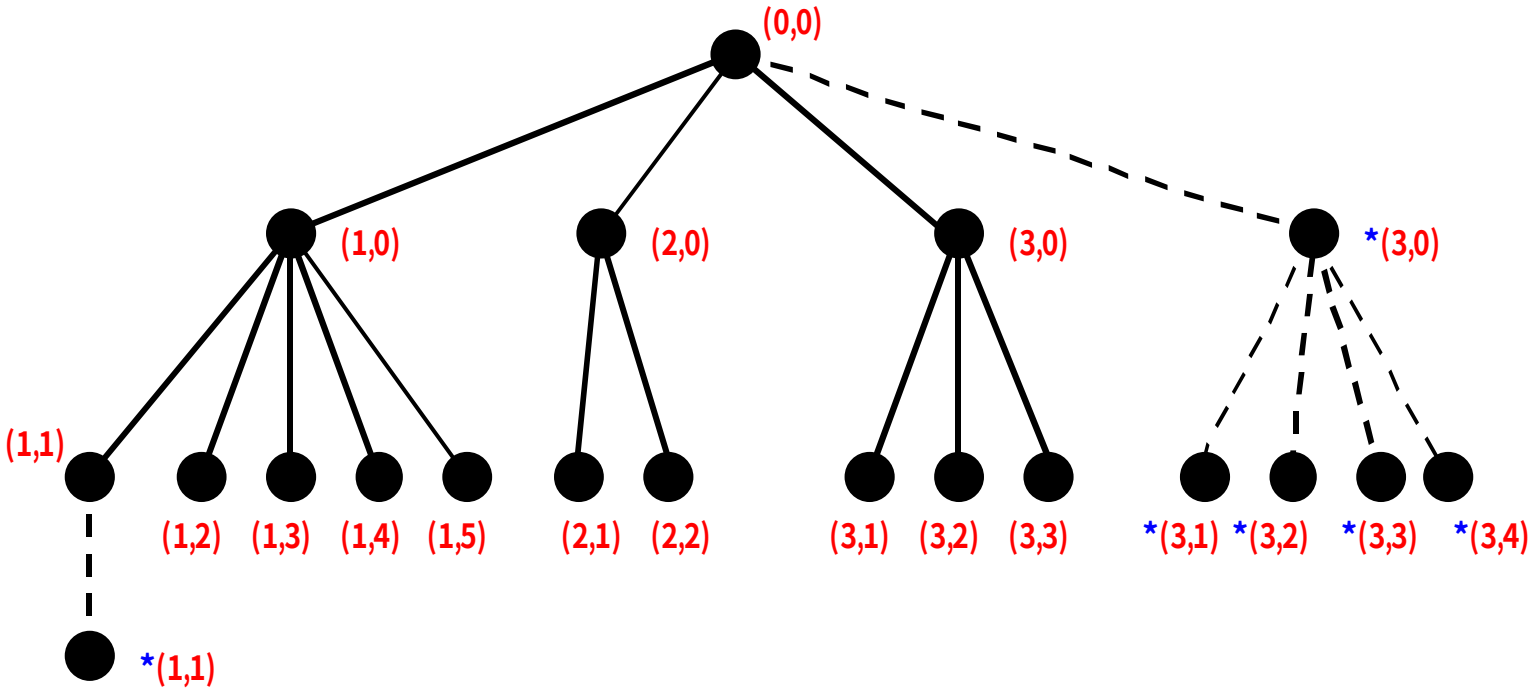
# 階層構造を識別可能な番号付け

- 数字の組による節点の位置表現
- コード番号による節点の位置表現
- オーバフローの表現
  - レベルのオーバフロー
  - 子の数のオーバフロー

# 数字の組による節点の位置表現

- 高さ $h$ までの木について，$h$ 組の数字からなる番号を付ける．
- レベル$n$ にある節点の，下から$h$-$n$ 個は0とする．
- レベル$n$ にある節点$s$ の親を$p$ とするとき，$s$ の位置を表す数字の組の先頭の$n$ -1組は$p$ を受け継ぐ．
- $n$ 番目の数字は，$p$ の子を左から数えた番号とする．

# 数字の組による節点の位置表現 - 例 -



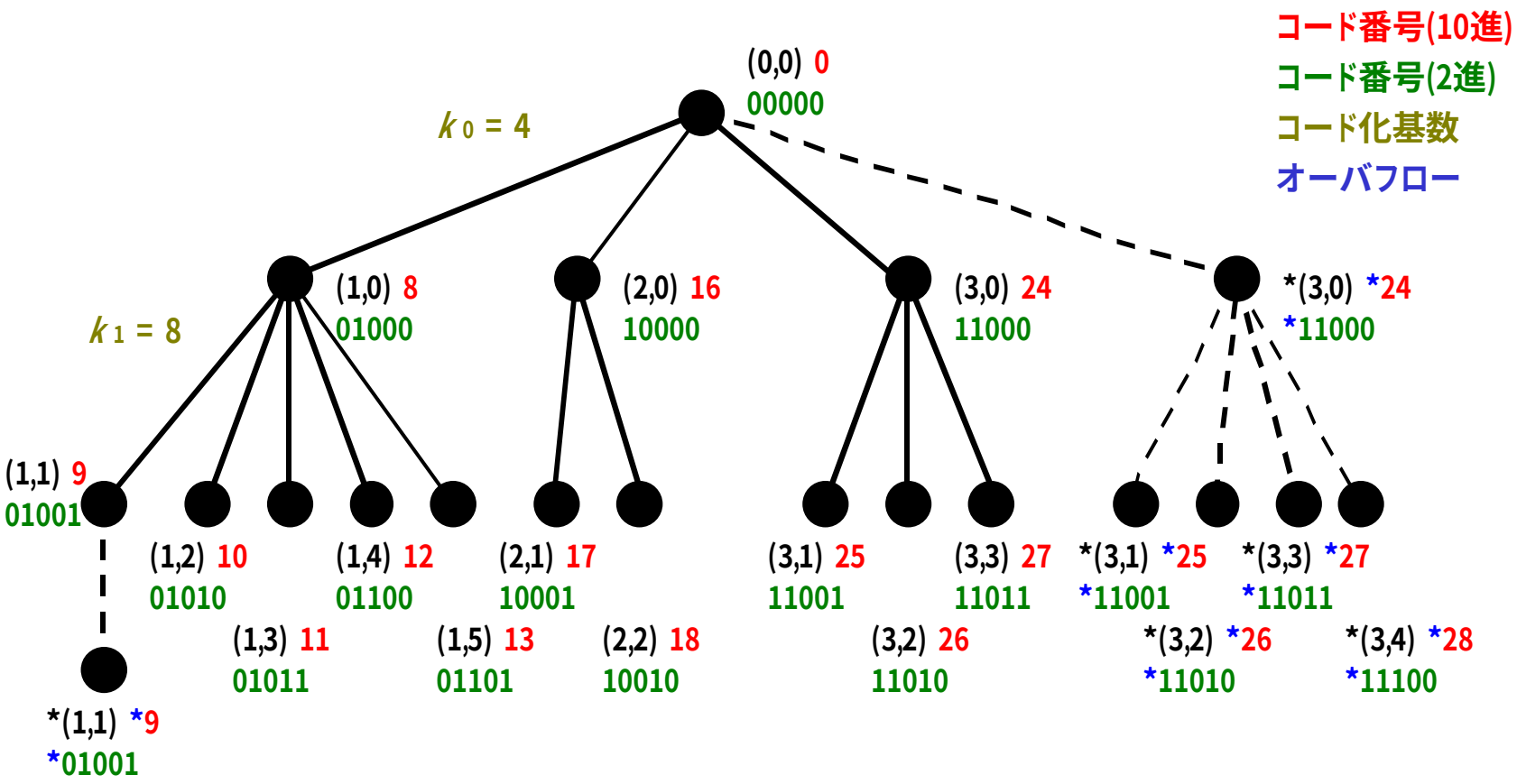
# コード番号による節点の位置表現

- レベルごとに異なる基数によるコーディングを行う．
- $k_i$：レベル$i$ $(0 \leqq i \leqq h-1)$の基数
- 番号を$m$ bitでコーディングする．オーバフローフラグに1bitを使用するため，実質$m-1$ bit $(m-1 \leqq \log_2(k_0 k_1 k_2 \ldots k_{h-1}))$
- 数字の組による位置表現$(a_1, \ldots, a_n, 0, \ldots, 0)$のコーディング式

$$((\ldots(a_1 k_1 + a_2)k_2 + \ldots + a_{n-1})k_{n-1} + a_n)k_n \ldots k_{h-1} \qquad (1)$$

但し，節点はレベル$n$ $(\leqq h)$にある．

# コード番号による節点の位置表現　- 例 -



# オーバフローの表現

- **レベルのオーバフロー**

  $h$ を超える深いレベルの節点は，オーバフローフラグを立て，その親と同じコードを割り当てる．

- **子の数のオーバフロー**

  レベル $i$ の子の数が $k_i$-1を超える場合，$k_i$ 番目以降の子のコードは、オーバフローフラグを立て，$k_i$-1番目の子と同じコードを付ける．その子の子孫についてもオーバフローフラグを立て，$k_i$-1番目の対応する位置に子孫がある場合と同じコードを割り当てる．

# 位置関係の計算

- 節点に付けられた番号から，節点間の位置関係を知る方法．
- 簡単のため，レベル $i$ の基数が2のべき乗で，$k_i = 2^{w_i}$ と表される場合について説明．
- 絶対位置の復元
- 節点間の関係
  - 最も近い共通先祖(LCA）
  - 先祖子孫関係にある節点間の距離
  - 兄弟および従兄弟関係

# 絶対位置の復元

- 節点に与えられた番号の2進数表現の下位から，bit幅 $w_{h-1}, w_{h-2}, \ldots, w_1, w_0$ を順に取り出す．
- 数字に変換したものをそれぞれ，$r_h, r_{h-1}, \ldots, r_2, r_1$とする．
- それらを逆順に並べて組にした$(r_1, r_2, \ldots, r_{h-1}, r_h)$は，その節点の数字の組による位置表現になる．

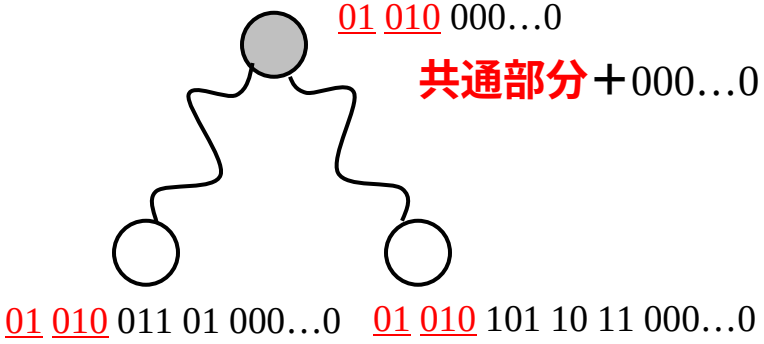
# 節点間の関係

- 節点に付けられた番号は，木のpre-order探索順で昇順に並ん(extended pre-order)でおり，且つ任意の節点の子孫の最大番号も簡単な計算で分かる．
- 2節点が，左，右，先祖あるいは子孫のいずれの関係であるかは，これら節点に付けられた番号だけで簡単に計算できる．

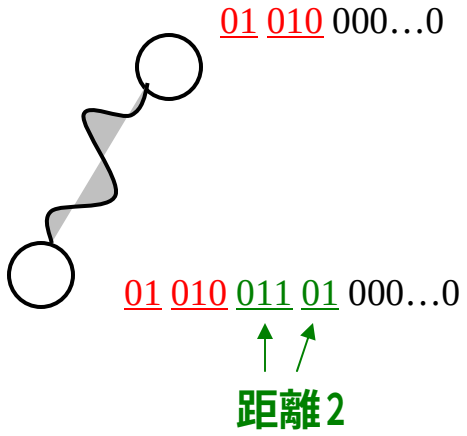
- 最も近い共通先祖(LCA)
- 先祖子孫関係にある節点間の距離
- 兄弟および従兄弟関係

# 節点間の関係の計算 ($k_0=2^2, k_1=2^3, k_2=2^3, k_3=2^2, \ldots$ の場合)

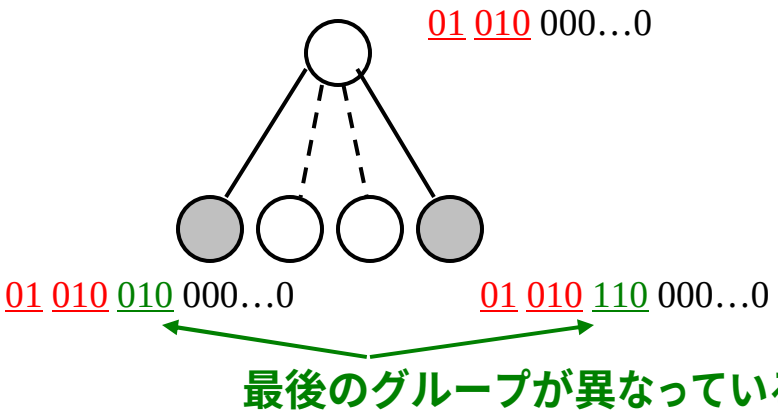
・最も近い共通先祖(LCA )



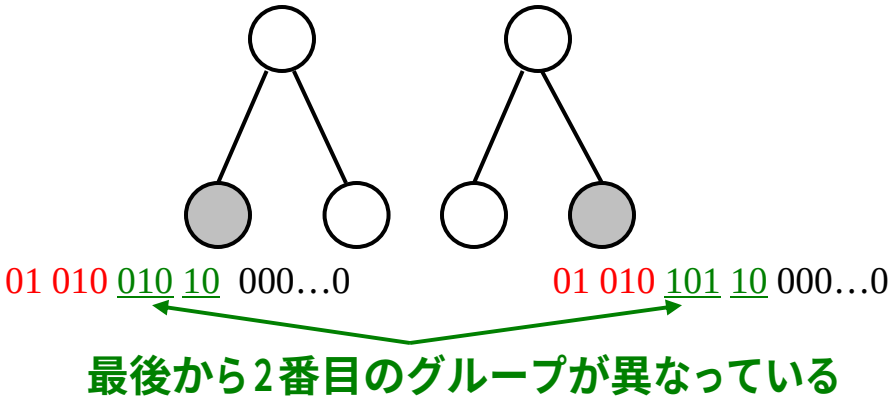

・先祖子孫関係にある節点間の距離




・兄弟関係




・従兄弟関係



# まとめ

- XML-DBにおける経路式質問を効率的に処理するため，索引に文書の階層構造を反映した番号を格納する．
- XML文書を表す木節点への新しい番号付け手法を提案した．
- これを用い，節点間の左右あるいは先祖子孫関係だけでなく，兄弟あるいは親子関係などのより詳細な情報を知ることができた．
- 各レベルの基数の選び方の詳細，番号のオーバフローの具体的処理方法，更新に伴う番号の付け替えなどの課題がある．