Query Optimization: Exercise

Session 11

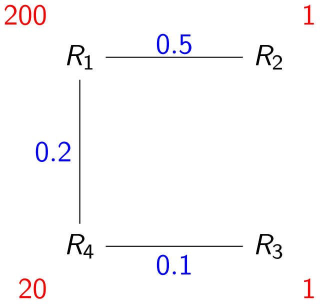
Bernhard Radke

January 15, 2018

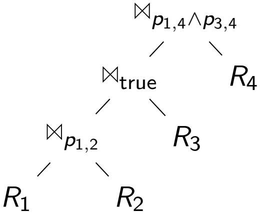
Order-Preserving Joins

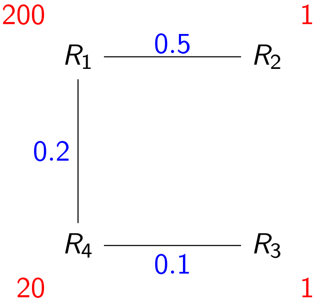
Consider the following *sequence* of relations R_1, R_2, R_3, R_4 with cardinalities $|R_1| = 200$, $|R_2| = 1$, $|R_3| = 1$, $|R_4| = 20$ and join selectivities $f_{1,2} = 0.5$, $f_{1,4} = 0.2$, $f_{3,4} = 0.1$.

Give the fully-parenthesized, optimal join-expression that abides by this order. Use C_{out} as a cost function.



Let's start off with a cost analysis of the left-deep tree:





$$C_{out} = 100 + 100 + 40 = 240$$

OrderPreservingJoins($R = \{R_1, \ldots, R_n\}$,P)
Input: a set of relations to be joined and a set of predicates
Output: fills p, s, c, t
for each $1 \leq i \leq n$ {
  $p[i,i]$ =predicates from P applicable to R_i
  $P = P \setminus p[i,i]$
  $s[i,i]$ =statistics for $\sigma_{p[i,i]}(R_i)$
  $c[i,i]$ =costs for $\sigma_{p[i,i]}(R_i)$
}

predicates p

$\emptyset$			
	$\emptyset$		
		$\emptyset$	
			$\emptyset$

statistics s

200			
	1		
		1	
			20

costs c

0			
	0		
		0	
			0

split points t

for each $2 \leq l \leq 4$ **ascending** (in text: $2 \leq l \leq n$)
  for each $1 \leq i \leq 5 - l$ (in text: $1 \leq i \leq n - l + 1$)
    $j = i + l - 1$
    $p[i,j]$=predicates from P applicable to $R_i, \ldots, R_j$
    $P = P \setminus p[i,j]$
    $s[i,j]$=statistics derived from $s[i,j-1]$ and $s[j,j]$ including $p[i,j]$
    $c[i,j]$=∞
    for each $i \leq k < j$
      $q = c[i,k] + c[k+1,j]$+costs for $s[i,k]$ and $s[k+1,j]$ and $p[i,j]$
      if $q < c[i,j]$
        c[i,j]=q
        t[i,j]=k

predicates p

$\emptyset$			
	$\emptyset$		
		$\emptyset$	
			$\emptyset$

statistics s

200			
	1		
		1	
			20

costs c

0			
	0		
		0	
			0

split points t

for each $2 \leq l \leq 4$ **ascending** (in text: $2 \leq l \leq n$)
  for each $1 \leq i \leq 5 - l$ (in text: $1 \leq i \leq n - l + 1$)
    $j = i + l - 1$
    $p[i,j]$=predicates from P applicable to $R_i, \ldots, R_j$
    $P = P \setminus p[i,j]$
    $s[i,j]$=statistics derived from $s[i, j-1]$ and $s[j,j]$ including $p[i,j]$
    $c[i,j]$=∞
    for each $i \leq k < j$
      $q = c[i,k] + c[k+1,j]$+costs for $s[i,k]$ and $s[k+1,j]$ and $p[i,j]$
      if $q < c[i,j]$
        c[i,j]=q
        t[i,j]=k

predicates p

$\emptyset$	$p_{1,2}$		
	$\emptyset$		
		$\emptyset$	
			$\emptyset$

statistics s

200	100		
	1		
		1	
			20

costs c

0	100		
	0		
		0	
			0

split points t

	1		

for each $2 \leq l \leq 4$ **ascending** (in text: $2 \leq l \leq n$)
  for each $1 \leq i \leq 5 - l$ (in text: $1 \leq i \leq n - l + 1$)
    $j = i + l - 1$
    $p[i,j]$=predicates from P applicable to $R_i, \ldots, R_j$
    $P = P \setminus p[i,j]$
    $s[i,j]$=statistics derived from $s[i,j-1]$ and $s[j,j]$ including $p[i,j]$
    $c[i,j]$=∞
    for each $i \leq k < j$
      $q = c[i,k] + c[k+1,j]$+costs for $s[i,k]$ and $s[k+1,j]$ and $p[i,j]$
      if $q < c[i,j]$
        c[i,j]=q
        t[i,j]=k

predicates p

$\emptyset$	$p_{1,2}$		
	$\emptyset$	$\emptyset$	
		$\emptyset$	
			$\emptyset$

statistics s

200	100		
	1	1	
		1	
			20

costs c

0	100		
	0	1	
		0	
			0

split points t

	1		
		2	

for each $2 \leq l \leq 4$ **ascending** (in text: $2 \leq l \leq n$)
  for each $1 \leq i \leq 5 - l$ (in text: $1 \leq i \leq n - l + 1$)
    $j = i + l - 1$
    $p[i,j]$=predicates from P applicable to $R_i, \ldots, R_j$
    $P = P \setminus p[i,j]$
    $s[i,j]$=statistics derived from $s[i,j-1]$ and $s[j,j]$ including $p[i,j]$
    $c[i,j]$=∞
    for each $i \leq k < j$
      $q = c[i,k] + c[k+1,j]$+costs for $s[i,k]$ and $s[k+1,j]$ and $p[i,j]$
      if $q < c[i,j]$
        c[i,j]=q
        t[i,j]=k

predicates p

$\emptyset$	$p_{1,2}$		
	$\emptyset$	$\emptyset$	
		$\emptyset$	$p_{3,4}$
			$\emptyset$

statistics s

200	100		
	1	1	
		1	2
			20

costs c

0	100		
	0	1	
		0	2
			0

split points t

	1		
		2	
			3

The values of t are:

i/j	1	2	3	4
1		1	1	1
2			2	3
3				3
4				

ExtractPlan($R = \{R_1, \ldots, R_n\}$,t,p)
Input: a set of relations, arrays t and p
Output: a bushy join tree
return ExtractPlanRec(R,t,p,1,n)

ExtractPlanRec($R = \{R_1, \ldots, R_n\}$,t,p,i,j)
if $i < j$
  T_1 =ExtractPlanRec(R,t,p,i,$t[i,j]$)
  T_2 =ExtractPlanRec(R,t,p,$t[i,j]+1$,j)
  return $T_1 \bowtie^L_{p[i,j]} T_2$
else
  return $\sigma_{p[i,j]} R_i$

The values of t are:

i/j	1	2	3	4
1		1	1	1
2			2	3
3				3
4				

extract-subplan(. . . , i=1, j=4)
    extract-subplan(. . . , i=1, j=1)
    extract-subplan(. . . , i=2, j=4)
        extract-subplan(. . . , i=2, j=3)
            extract-subplan(. . . , i=2, j=2)
            extract-subplan(. . . , i=3, j=3)
        return $(R_2 \bowtie_{\text{true}} R_3)$
        extract-subplan(. . . , i=4, j=4)
    return $((R_2 \bowtie_{\text{true}} R_3) \bowtie_{p_{3,4}} R_4)$
return $(R_1 \bowtie_{p_{1,2} \wedge p_{1,4}} ((R_2 \bowtie_{\text{true}} R_3) \bowtie_{p_{3,4}} R_4))$

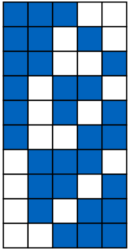
The total cost of this plan is $c[1,4] = 43$.

Combinatorics 101

Given a set of n elements, how many distinct k-element subsets can be formed?

$$\binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

Example: Choose 3 out of 5: $\binom{5}{3} = \frac{5!}{2! \cdot 3!} = \frac{120}{2 \cdot 6} = 10$

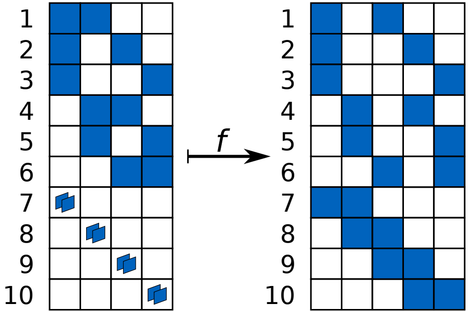


Homework

$$\binom{n}{k} = \binom{n}{n-k}$$

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$$

- Now *with replacement*: How many distinct *multisets* exist chosing k from n?
  As many as there are distinct sets chosing k from $n+k-1$!
- Bijection between multisets and sets. From multiset to set:
  $f : (x_1, x_2, \ldots, x_k) \mapsto (x_1 + 0, x_2 + 1, \ldots, x_k + (k-1))$
- Example: Choose 2 from 4
  - $\#$ sets: $\binom{4}{2}$
  - $\#$ multisets: $\binom{4+2-1}{2}$



- Slides and exercises: db.in.tum.de/teaching/ws1718/queryopt
- Send any questions, comments, solutions to exercises etc. to radke@in.tum.de
- Exercise due: 9 AM, January 22