Finding Complex Patterns over Data Streams

Leila Kaghazian Dennis McLeod Reza Sadri

Integrated Media Systems Center, University of Southern California Procom Technology

kaghazia@usc.edu *mcleod@usc.edu* *sadri@procom.com*

Introduction

Many Applications

- Querying purchase patterns for marketing
- Stock market analysis
- Studying meteorological data

What's needed:

- Expressive query language for finding complex patterns over data streams
- Efficient and scalable implementation: "Query *Optimization"*

SQL-TS

A query language for finding complex patterns in sequences

- Minimal extension of SQL—only the **from** clause affected
- A new Query optimization technique based on extensions of the Knuth, Morris & Pratt (KMP) string-search algorithm

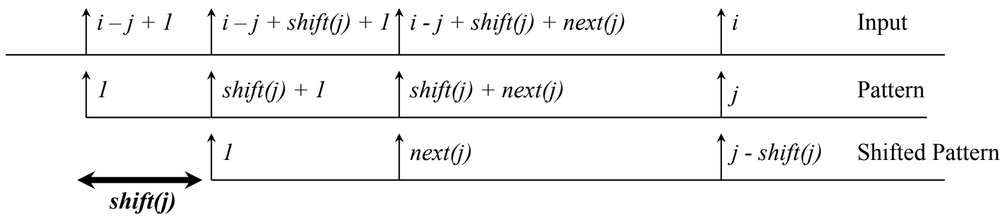
Optimized String Search: KMP

After failing, use the information acquired so to:

- backtrack to ***shift(j)***, rather than *i+1, and*
- only check pattern values after ***next(j)***

shift and **next**

- Success for first j-1 elements of pattern. Failure for jth element (when input is at i)
- Any shift less than shift(j) is guaranteed to lead to failure,
- Match elements in the pattern starting at next(j)



Matrices θ and ϕ

Input tested on pj is now tested against pk

pj succeeded: $\theta_{jk} = \begin{cases} 1 & if\ p_j \Rightarrow p_k \quad and \quad p_j \neq False \\ 0 & if\ p_j \Rightarrow \neg p_k \\ U & otherwise \end{cases}$

pj failed: $\phi_{jk} = \begin{cases} 1 & if\ \neg p_j \Rightarrow p_k \quad and \quad p_j \neq True \\ 0 & if\ \neg p_j \Rightarrow \neg p_k \\ U & otherwise \end{cases}$

Combing values of these lower triangular matrices (j ≥ k),
We derive the values of next(j) and shift (j)

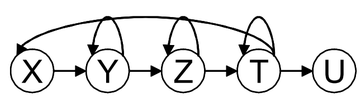
Nested Star Patterns vs. Star Patterns

```
SELECT  X.first.magnitude,
   X.first.time,
   U.previous.magnitude,
   U.previous.time
FROM earthquake
   CLUSTER BY region
   SEQUENCE BY time
   AS (*(X,*Y,*Z,*T),U)
WHERE
   X.region="Los Angeles"
   AND 1.8<X.magnitude
   AND X.magnitude<2.5
   AND Y. magnitude <
        Y.previous. magnitude
   AND .99.*Z.previous.magnitude
        < Z. magnitude
   AND Z.magnitude <
        1.1*Z.previous. magnitude
   AND T.magnitude >
        1.1*T.previous. magnitude
   AND U.magnitude > 3.5
```

```
SELECT X.NEXT.date,
   X.NEXT.price,
      S.previous.date,
   S.previous.price
FROM quote
   CLUSTER BY name,
   SEQUENCE BY date
   AS (*X, Y, *Z, *T, U, *V, S)
WHERE
   X.name='IBM'
   AND X.price > X.previous.price
   AND 30 < Y.price  AND Y.price
    < 40
   AND Z.price <  Z.previous.price
   AND T.price > T.previous.price
   AND 35 < U.price
   AND U.price < 40
   AND V.price < V.previous.price
   AND S.price < 30
```

Handling Nested Star Patterns

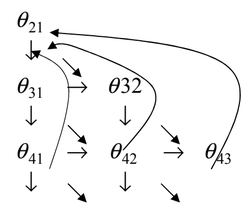
State model



Adjacency Matrix

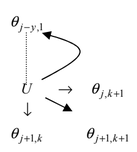
$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Same input, Transitions on Original Pattern vs.
Transitions on Pattern after the index set back j-k.



Example

Element j is the last element of the nested star sub-pattern and is a star predicate, element k is a star predicate and θ_{jk} is U



Directed graph produced by possible transitions between pattern elements will be called the Implication Graph for pattern sequence P.

$$G_p = \begin{bmatrix} - & & & & \\ U & - & & & \\ 0 & U & - & & \\ 1 & U & 0 & - & \\ 0 & U & 1 & 0 & - \\ U & 0 & U & U & U & - \end{bmatrix}$$

Implication Graph

Calculating next and shift

shift(j) : $\min\{s \mid \exists t$ where there is path from $\theta_{s+1,1}$ to $\varphi_{j,t}$ in $G_P^j\}$

next(j) :

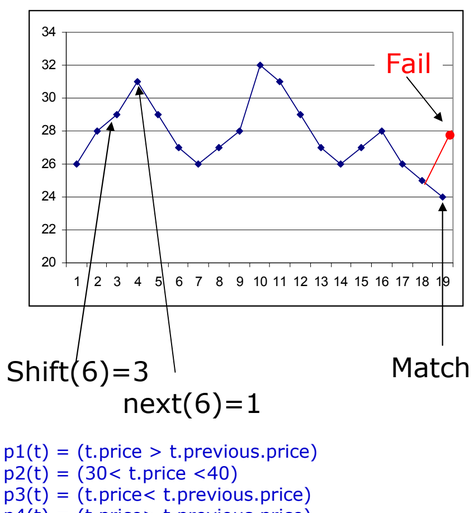
If shift(j)<j-1 then

next(j) = $\min\{s \mid \exists a, t \text{ s.t. } \theta_{as} = U$ in the path from $\theta_{shift(j)+1,a}$ to ϕ_{jt} in G_P^j and there is no fork in the path from $\theta_{shift(j)+1,1}$ to $\theta_{as}\}$

else next(j)= j-shift(j)

shift(j) and next(j) Example

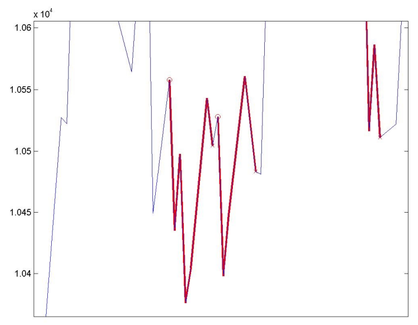
```
SELECT X.next.date, X.next.price,
 S.previous.date, S.previous.price
FROM quote
 CLUSTER BY name,
 SEQUENCE BY date
 AS (*X,*(Y,*Z,*T),*U,V)
WHERE
 X.name="Intel"
 AND X.price>X.previous.price
 AND 30<Y.price
 AND Y.price<40
 AND Z.price <Z.previous.price
 AND T.price > T.previous.price
 AND U.price<U.previous.price
 AND V.price<25
```



p1(t) = (t.price > t.previous.price)
p2(t) = (30< t.price <40)
p3(t) = (t.price< t.previous.price)
p4(t) = (t.price> t.previous.price)
p5(t) = (t.price < t.previous. price)
p6(t) = (t.price < 25)

Experimental result

Pattern of consecutive interval of period of high fluctuation (more than 1% up and down) followed by a period of steady raise.



Improving and Extending OPS*

Buffer Size
Data Compression
Multi Dimensional Data Streams
Pattern Dependent Search