Searching Correlated Objects in a Long Sequence

Ken C. K. Lee[1] Wang-Chien Lee[1] Donna Peuquet[1] Baihua Zheng[2]

cklee@cse.psu.edu wlee@cse.psu.edu djp11@psu.edu bhzheng@smu.edu.sg

[1] Pennsylvania State University, University Park, PA16802, USA

[2] Singapore Management University, Singapore

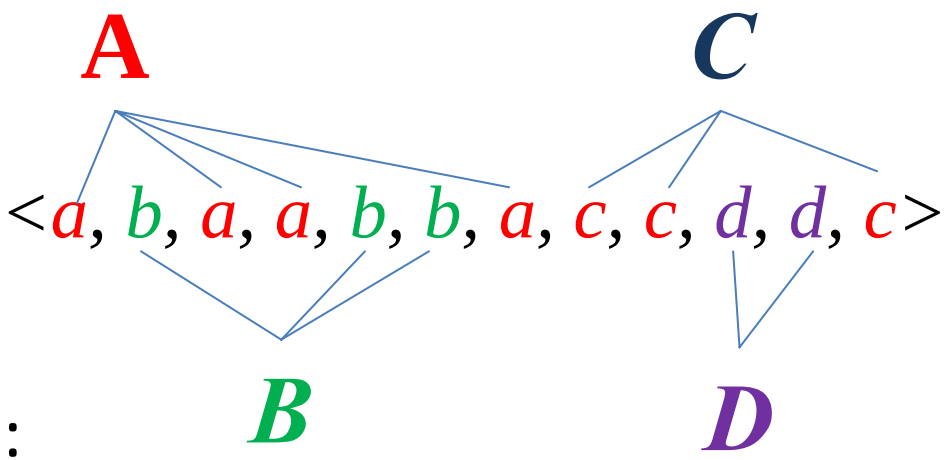
SSDBM, Hong Kong. Jul 9-10 2008

Outline

- Overview of Correlation Query
- Related Work
- Motivation and Definition
- Challenges
- Solutions
  - Non-Index Approaches
    - Multi-Scan Algorithm (MSA)
    - One-Scan Algorithm (OSA)
  - Index-Based Approach (IBA)
- Evaluations
- Variants of Correlation Query
- Conclusion

Overview of Correlated Query

- Sequence
  - An ordered list of objects (categorized by their attributes)
  - A working example:



Object sets:

A={a_1, a_3, a_4, a_7}, B={b_2, b_5, b_6}, C={c_8,c_9,c_{12}}, D={d_{10},d_{11}}

- Correlation Query:
  - Given a sequence of objects, find pairs of correlated object sets which has many **objects closely located** in the sequence.

Related Work

- **Statistics**: how are the *values* of one variable (education) related to those of another (income)?
- **Database**: how are the *occurrences* of object related to those of another (e.g. in same transactions)?

If x and y is highly correlated, f_{xy} should be high (relative to N)

	y	$\bar{y}$	
x	f_{xy}	$f_{x\bar{y}}$	f_x
$\bar{x}$	$f_{\bar{x}y}$	$f_{\bar{x}\bar{y}}$	$f_{\bar{x}}$
	f_y	$f_{\bar{y}}$	N

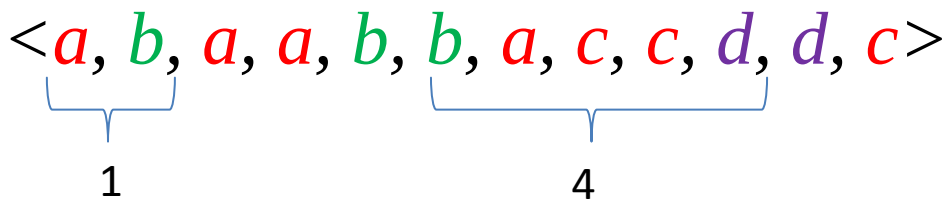
Contingency table

Motivation

- Applications
  - Finding *products likely to be chosen* by customers, based on transaction logs.
  - *Event causality* detection based on event log to determine what events are likely to happen after some events.
  - In documents, identify *word phrases* (composition of words) that are often used.

Definition

- Object Closeness
  - Objects Distance
    - Based on difference between sequence positions



- Correlation Coefficient
  - How many pairs of closely located objects?
  - Based on cosine coefficient

$$\Phi_w(X,Y) = \frac{|XY|_w}{\sqrt{|X|\,|Y|}}$$

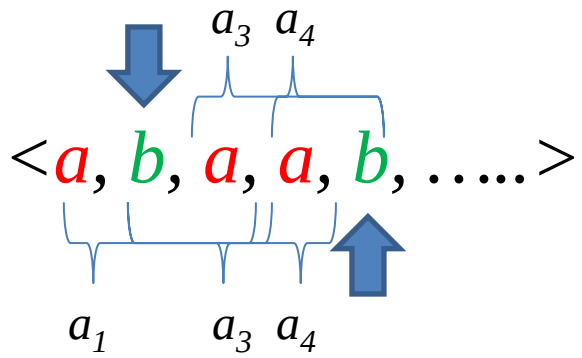
No. of closely object pairs (determined by *w*)

No. of objects belonging to X

No. of objects belonging to Y

Challenges

- Redundant Count Problem
  - Let $w = 2$



$|AB|_2 = 5$!? (This is incorrect)

  - Close object pairs must be disjoint
- Correlation coefficient $\Phi_w(X,Y)$ is subject to w.

Query Definition

- Correlation Query
  - Given a sequence ***S***, a set of predefined object sets, ***O***, and two query parameters, distance bound (***w***) and correlation threshold (***t***),

    a correlation query ***Q(S,w,t)*** returns all pairs of object sets $(X,Y) \in OxO$ such that $\Phi_w(X,Y) > t$.
- Example

  <*a*, *b*, *a*, *a*, *b*, *b*, *a*, *c*, *c*, *d*, *d*, *c*>

  A={a_1, a_3, a_4, a_7},
  B={b_2, b_5, b_6},
  C={c_8,c_9,c_{12}},
  D={d_{10},d_{11}}

w=2

XY	$\|X\|$	$\|Y\|$	$\|XY\|_\omega$	$\phi_\omega(X,Y)$
AB	4	3	3	0.87
AC	4	3	1	0.29
AD	4	2	0	0.00
BC	3	3	1	0.33
BD	3	2	0	0.00
CD	3	2	2	0.82

Solutions

- Scan-Based Approaches (Scan S to determine $|X|$, $|Y|$, $|XY|_w$)
  - Multi-Scan Algorithm (MSA) – baseline approach
    - determine one X Y pair each time
  - One-Scan Algorithm:
    - determine all X Y pairs in one scan
- Index-Based Approach
  - Index-Based Algorithm (IBA)
    - Index the objects and their position based on object set
    - Determine possible X Y pairs whose $|XY|_w$ are high.

Multi-Scan Algorithm (MSA)

- Scan S for each X Y pair
- Three counters c_X, c_Y and c_{XY} (initialized to zeroes)
- Sliding window W (len: w)
- Example:

always matching the oldest entry in W

object	W	matched	c_A	c_B	c_{AB}
$\langle$init$\rangle$	$(\perp,\perp)$	-	0	0	0
a_1	$(\perp,a_1)$	no	1	0	0
b_2	$(\sout{a_1},\sout{b_2})$	$\langle a_1, b_2\rangle$	1	1	1
a_3	$(\sout{b_2},a_3)$	no	2	1	1
a_4	(a_3,a_4)	no	3	1	1
b_5	$(a_4,\sout{b_5})$	$\langle a_3, b_5\rangle$	3	2	2
b_6	$(\sout{b_5},\sout{b_6})$	$\langle a_4, b_6\rangle$	3	3	3
a_7	$(\sout{b_6},a_7)$	no	4	3	3
c_8	$(a_7,\perp)$	no	4	3	3
c_9	$(\perp,\perp)$	no	4	3	3
d_{10}	$(\perp,\perp)$	no	4	3	3
d_{11}	$(\perp,\perp)$	no	4	3	3
c_{12}	$(\perp,\perp)$	no	4	3	3

$$\phi(A,B) = c_{AB}/\sqrt{c_A \times c_B} = 3/\sqrt{4\times 3} = 0.87$$

- Time complexity: $O(w\ |O|^2\ |S|)$

One-Scan Algorithm (OSA)

- Scan the sequence of all X Y pairs in one pass
- Maintain counters for each object set and counters for object set combinations
- Sliding Window W (len: w)
  - Entry format: (x : {y})
- Example

Each entry is associated with matched objects

exam	W	c_{AB}	c_{AC}	c_{BC}
$\langle$init$\rangle$	$(\perp, \perp)$	0	0	0
a_1	$(\perp, a_1{:}\{\})$	0	0	0
b_2	$(a_1{:}\{b_2\}, b_2{:}\{a_1\})$	1	0	0
a_3	$(b_2{:}\{a_1\}, a_3{:}\{\})$	1	0	0
a_4	$(a_3{:}\{\}, a_4{:}\{\})$	1	0	0
b_5	$(a_4{:}\{\}, b_5{:}\{a_3\})$	2	0	0
b_6	$(b_5{:}\{a_3\}, b_6{:}\{a_4\})$	3	0	0
a_7	$(b_6{:}\{a_4\}, a_7{:}\{\})$	3	0	0
c_8	$(a_7{:}\{c_8\}, c_8{:}\{a_7, b_6\})$	3	1	1
c_9	$(c_8{:}\{a_5, b_6\}, c_9{:}\{\})$	3	1	1
d_{10}	$(c_9{:}\{\}, d_{10})$	3	1	1
d_{11}	(d_{10}, d_{11})	3	1	1
c_{12}	$(d_{11}, c_{12} : \{\})$	3	1	1

- Time complexity: $\boldsymbol{O}(w\ |S|)$

Index-Based Algorithm (IBA)

- Index object positions for each object set
  - For example, $<a, b, a, a, b, b, a, c, c, d, d, c>$

    A=<1,3,4,7>

    B=<2,5,6>

    C=<8,9,12>

    D=<10,11>
  - Merge-like matching function

Similar to MSA, but it skips unrelated objects in the rest of the sequence.

A	C	W	c_{AC}
$\langle init \rangle$	$\langle init \rangle$	$(\perp,\perp)$	0
$\underline{a_1}$	c_8	$(\perp,a_1)$	0
$\underline{a_3}$	c_8	$(\perp,a_3)$	0
$\underline{a_4}$	c_8	(a_3,a_4)	0
$\underline{a_7}$	c_8	$(\perp,a_7)$	0
–	$\underline{c_8}$	$(\sout{a_7},\sout{c_8})$	1
–	$\underline{c_9}$	$(\sout{c_8},c_9)$	1
–	$\underline{c_{12}}$	(c_9,c_{12})	1

IBA Optimization Techniques

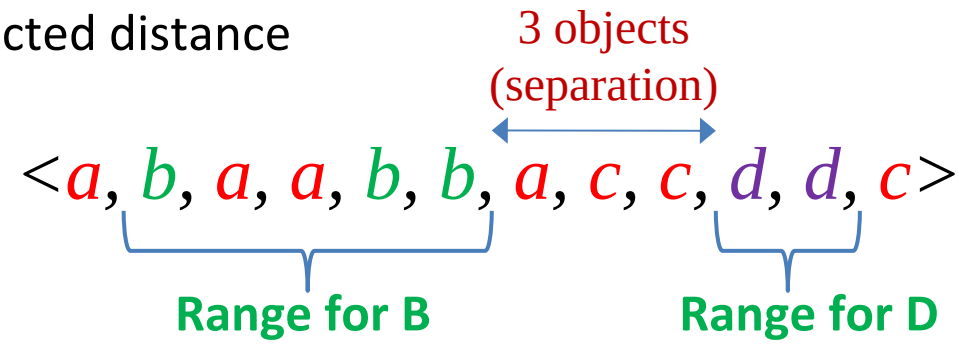
- Candidate Screening
- Group Matching
- Early Termination

IBA Optimization Techniques

- Candidate Screening
  - Estimation based on cardinalities

$$\Phi_w(X,Y) = \frac{|XY|_w}{\sqrt{|X|\,|Y|}} \qquad MAX\ \Phi(X,Y) = \frac{\min(|X|, |Y|)}{\sqrt{|X|\,|Y|}}$$

    - If $MAX\ \Phi(X,Y)$ is ***below*** the threshold, X Y are not correlated.
  - Estimation based on distribution
    - Based on range and expected distance



    - If ranges (extended with w) do ***not*** overlap, XY are not correlated.

IBA Optimization Techniques

- Assumption:
  - object separation can be modeled as normal distribution.
- We estimate the probability

  i.e., $P(-\omega \leq \delta_{X,Y} \leq \omega)$ $\quad$ $P(|\delta_{X,Y}| \leq \omega)$
- Based on Central Limit Theorem:

$$Z = \frac{(\mu_X - \mu_Y) - \delta_{X,Y}}{\sqrt{\sigma_X^2/|X| + \sigma_Y^2/|Y|}}$$

The probability p = $P(-\infty \leq Z \leq z_{upper}) - P(-\infty \leq Z \leq z_{lower})$

where $z_{lower} = \frac{(\mu_X - \mu_Y) - \omega}{\sqrt{\sigma_X^2/|X| + \sigma_Y^2/|Y|}}$ $z_{upper} = \frac{(\mu_X - \mu_Y) + \omega}{\sqrt{\sigma_X^2/|X| + \sigma_Y^2/|Y|}}$

Estimated correlation coefficient: $p \cdot \frac{\min(|X|,|Y|)}{\sqrt{|X| \cdot |Y|}}$

IBA Optimization Techniques

- Group Matching:
  - Rather than comparing two object sets each time, scan all possible pairs one pass (similar to OSA)
- Early Termination:
  - Maximum coefficient:

$$\frac{c_{XY}+\min(|X|-c_X+\omega_X,|Y|-c_Y+\omega_Y)}{\sqrt{|X||Y|}}$$

# of X objects examined

# of X objects in sliding window

# of paired objects found

# of X objects examined

# of X objects in sliding window

  - Minimum coefficient:

$$\frac{c_{XY}}{\sqrt{|X||Y|}}$$

Termination condition:

•Maximum coefficient < t, X Y should not be a part of the result.

•Minimum coefficient >= t, X Y is guaranteed to be a part of the result.

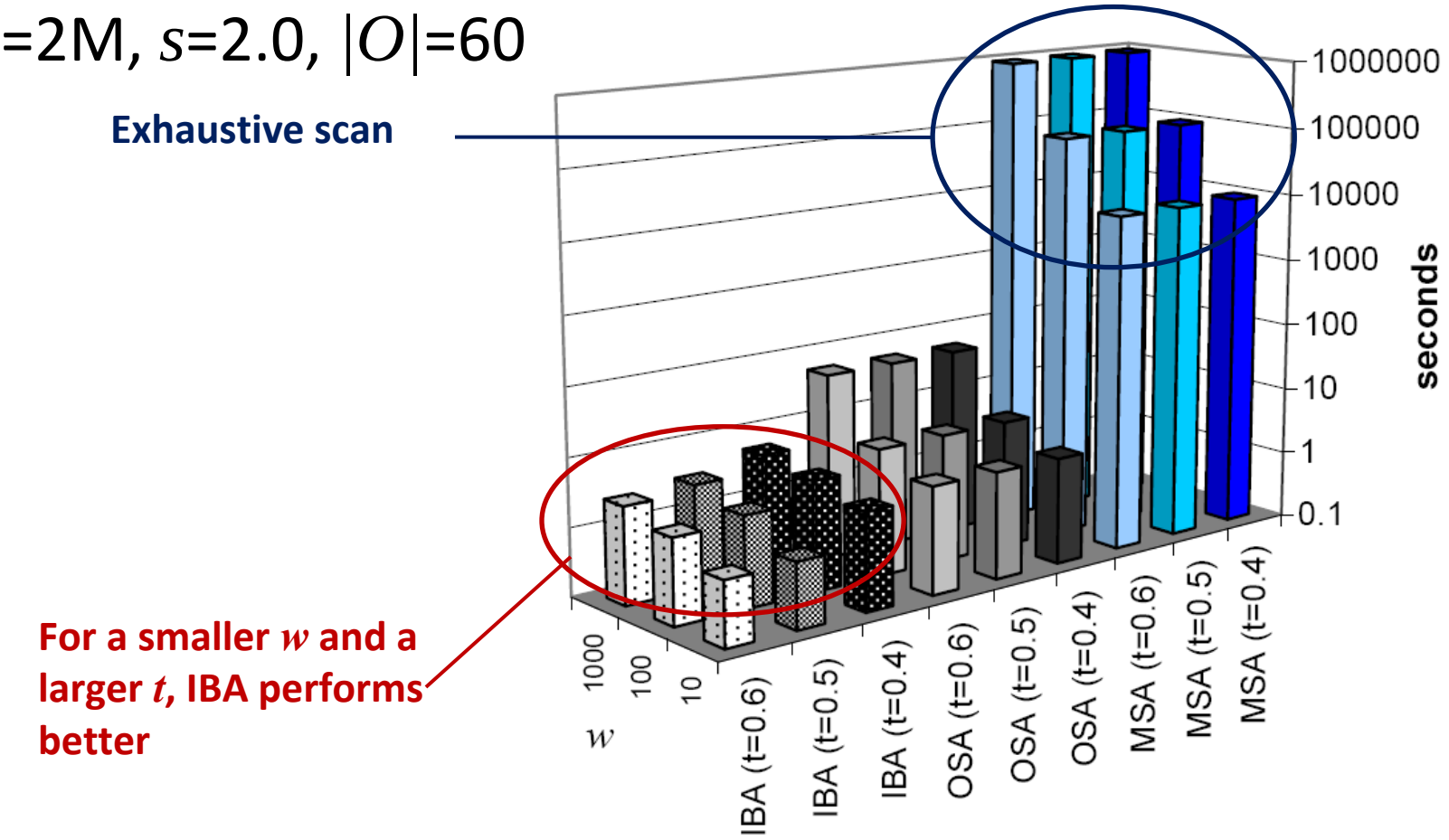
Performance Evaluation

- Query factors:
  - Object closeness (w): 10, 100 and 1000
  - Correlated Coefficient (t): 0.4, 0.5 and 0.6
- Datasets:
  - Synthetic datasets
    - Factors:
      - Zipf distribution skewness factor: 1.5 – 3.0 (default: 2.0)
      - Sequence length: 1M – 5M (default: 2M)
      - Number of object sets: 20 – 100 (default: 60)
  - Realistic datasets
    - EARTHQUAKE (http://earthquake.usgs.gov/region/neic)
    - APRS (http://aprs.net)
- Performance Metrics:
  - elapsed time
- Platform:
  - Linux Computer with 3.2GHz CPU

Impact of w and t

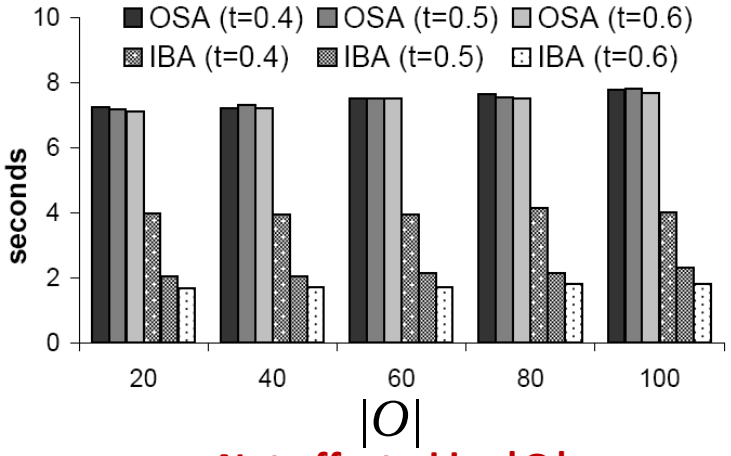
- Settings:

  $|S|$=2M, s=2.0, $|O|$=60

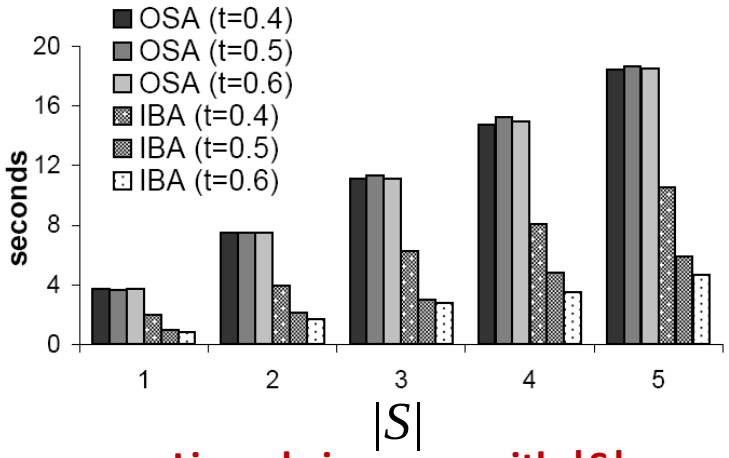


Impact of $|O|$ and $|S|$

- Impact of |O|
  - Fixed $|S|$, s and w at 2M, 2.0 and 100, respectively. Fixed $|O|$, s and w at 60, 2.0 and 100, respectively.
- Impact of |S|



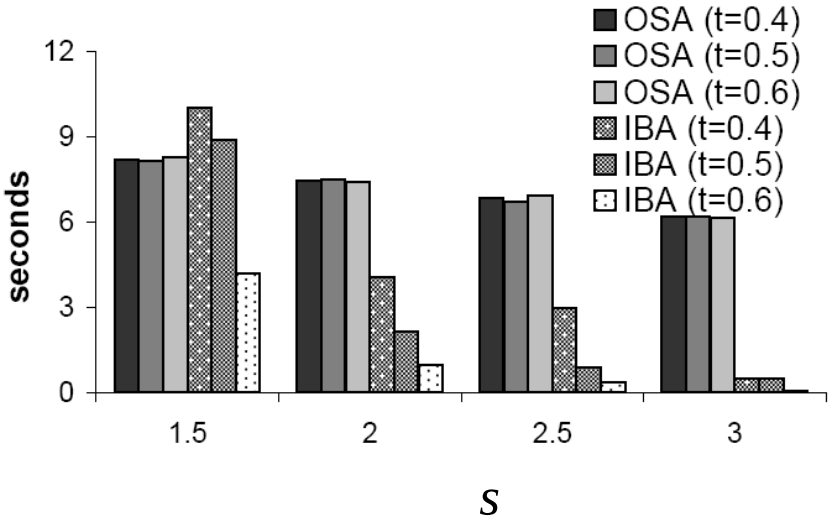
Not affected by |O|



Linearly increase with |S|

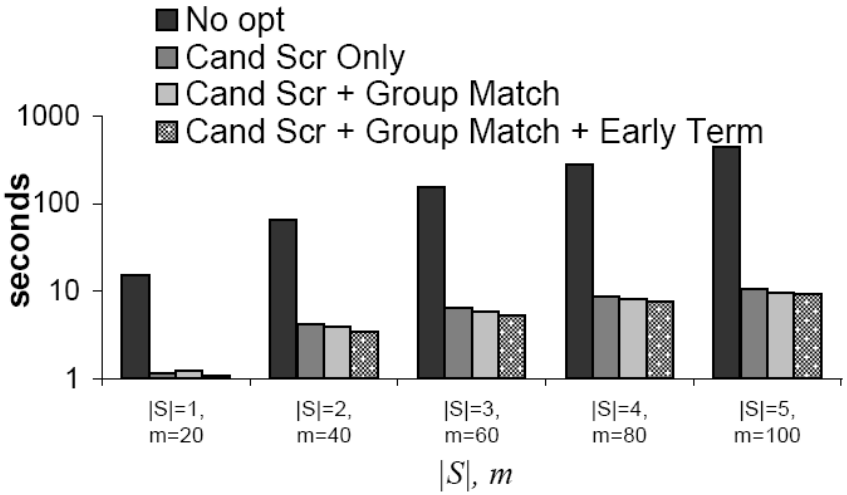
Impact of s, Effectiveness of optimization techniques

- Impact of s
  - Fixed $|O|$, $|S|$ and w at 60, 2M and 100, respectively.



The more skewed the sizes of object sets, the better IBA can perform

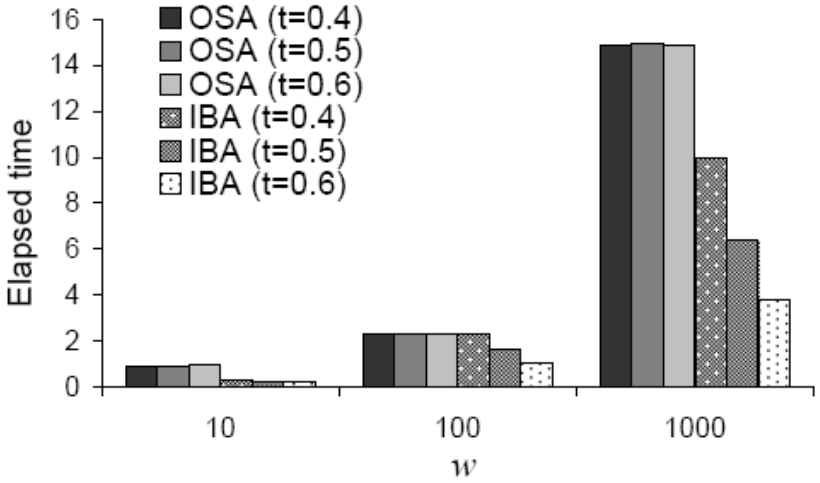
- Effectiveness of optimization techniques for IBA



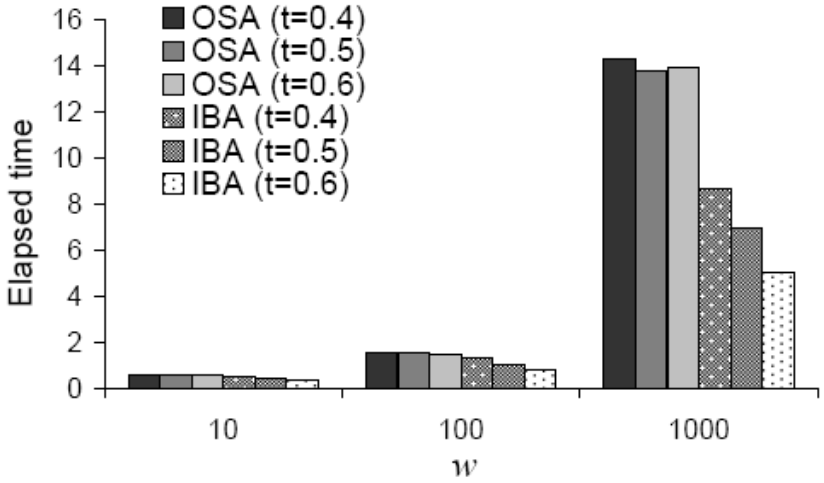
Candidate screening is the most effective

Evaluation on real datasets

- Earthquake



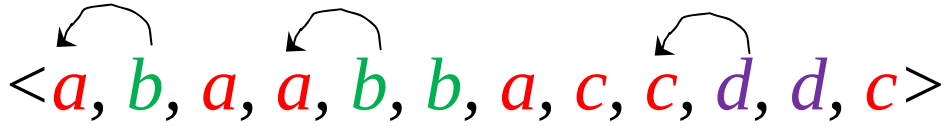
- APRS



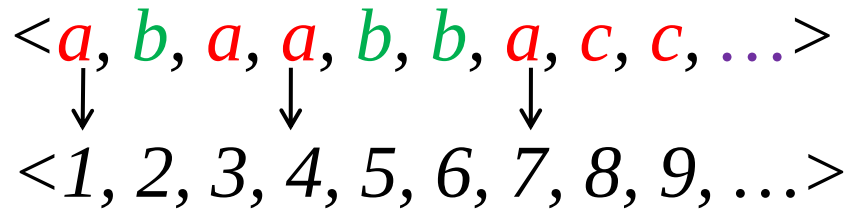
IBA outperform OSA

Variant Correlated Query

- Constrained Correlation Query
  - Limit the matching criteria



- Position Correlation Query



- Correlation Spectrum Query

w=2 $\Phi w(B,C)=0.33$

⋮

w=6 $\Phi w(B,C)=1.00$

Conclusion

- Introduced correlation query for a sequence
- Proposed search algorithms; MSA, OSA and IBA
- Experimented with synthetic and real datasets
- IBA generally performs good, especially for small w and large t and large variation of object set sizes
- Discussed correlation query variants

Thank you

Questions?