Relevance Queries for Interval Data

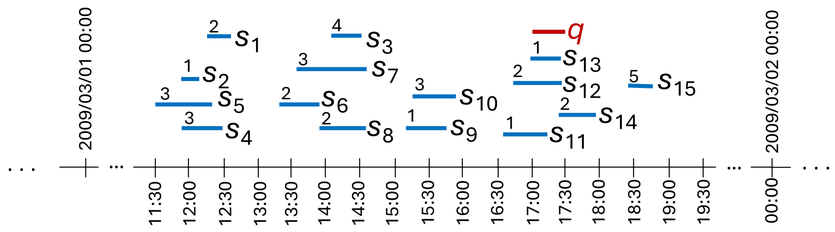
Panagiotis Bouros[1] and Nikos Mamoulis[2,3]

[1]Institute of Computer Science, Johannes Gutenberg University Mainz, Germany
[2]Department of Computer Science & Engineering, University of Ioannina, Greece
[3]Archimedes, Athena Research Center, Greece
bouros@uni-mainz.de, nikos@cse.uoi.gr

Motivation

Interval Data

- ❑ *Temporal* databases, *validity* intervals
- ❑ *Uncertain* data, *uncertainty* intervals
- ❑ *Anonymized* data, interval values on *sensitive* attributes



Range Queries

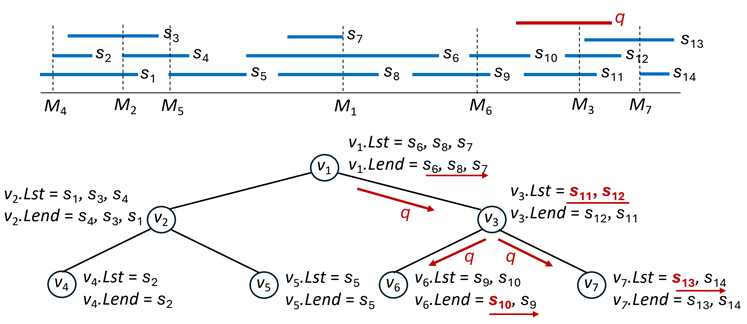
- ❑ *Fundamental* query operation
- ❑ Potentially *overwhelming* result *size*
- ❑ Need for *relevance-based search*

Interval Indexing

Interval tree

H. Edelsbrunner, Technical Report, TU Graz, Austria, 1980

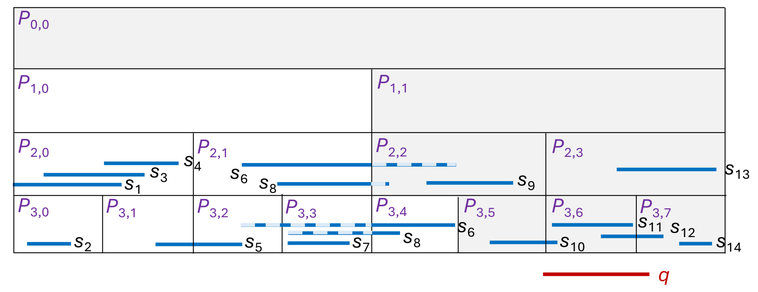
- ❑ *Binary search* tree with *O(n) space*
- ❑ Recursively partition space on *median*
- ❑ *Depth-first* traversal for queries



HINT

G. Christodoulou, P. Bouros and N. Mamoulis, ACM SIGMOD, 2022

- ❑ *Hierarchical*, *uniform* space partitioning
- ❑ Occupies *O(mn) space*, for *m+1* levels
- ❑ Store interval inside the *smallest set* of partitions from *all levels* covering it
- ❑ *Bottom-up* traversal for queries



Relevance-based Search

Absolute relevance $Rel_a(s,q) = |s \cap q|$

Data-relative relevance $Rel_{rd}(s,q) = \frac{|s \cap q|}{|s|}$

Relative relevance $Rel_r(s,q) = \frac{|s \cap q|}{|s \cup q|}$

Query-relative relevance $Rel_{rq}(s,q) = \frac{|s \cap q|}{|q|}$

$s \cap q = [\max\{s.start, q.start\}, \min\{s.end, q.end\}]$

$s \cup q = [\min\{s.start, q.start\}, \max\{s.end, q.end\}]$

$|s| = s.end - s.start$

Threshold-based search, *ϑRelQuery*

- ❑ All intervals with *relevance over ϑ*

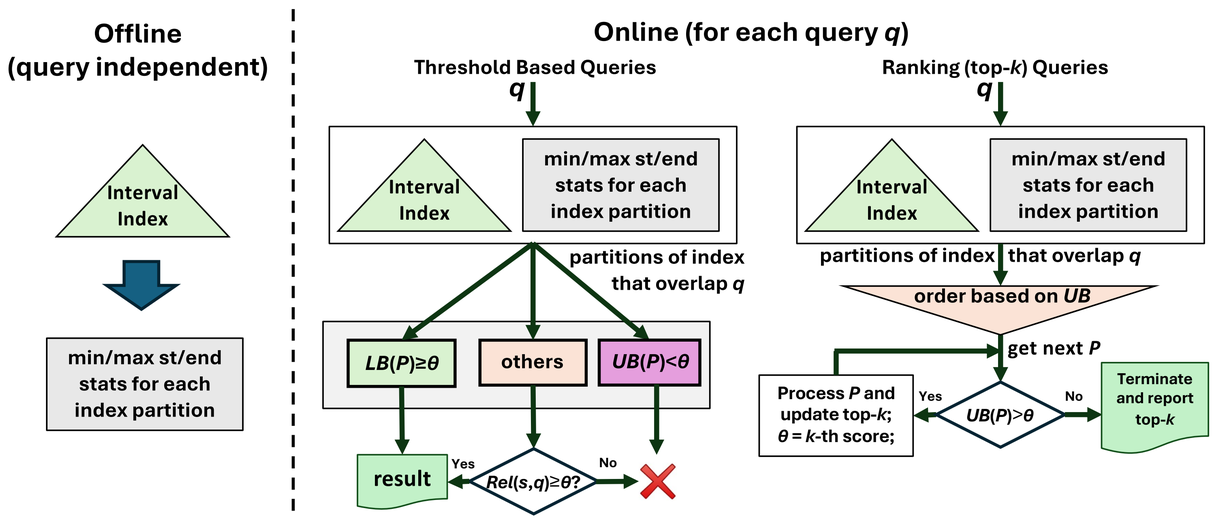
Ranking search, *kRelQuery*

- ❑ *k most relevant* intervals

Query Processing

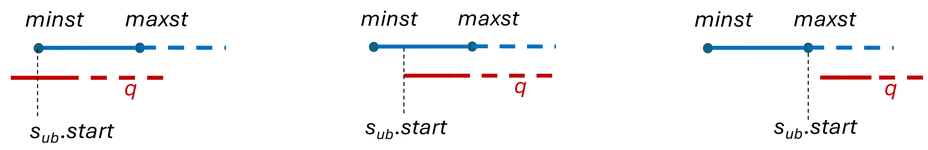
Unified Processing Framework

- ❑ *Applicable* to any interval indexing
- ❑ Requires *cheap-to-compute stats*, minimum and maximum endpoints



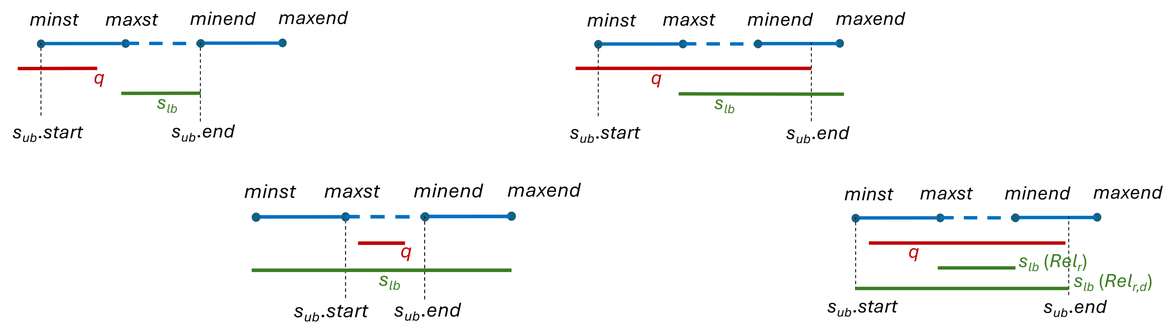
Upper Relevance Bound $UB(P) = Rel(s_{ub}, q)$

- ❑ *Shortest possible interval* s_{ub} *maximizing* absolute relevance $Rel(s_{ub}, q)$



Lower Relevance Bound $LB(P) = Rel(s_{lb}, q)$

- ❑ Interval s_{lb} *minimizing* absolute relevance
- ❑ While *maximizing* $|s_{lb} \cup q|$ for Rel_r and $|s_{lb}|$ for Rel_{rq}



Experiments

Setup

- ❑ Vary query interval extend, vary *ϑ* and *k*
- ❑ Query processing *with* or *without bounds*
- ❑ Also, for *k*RelQuery, *native* traversal or *best-first*

	BOOKS	WEBKIT	BTC	TAXIS
Cardinality	2,050,707	2,347,346	2,538,921	169,290,307
Size [MBs]	32	28	52	2,794
Domain	1 year	15 years	3 months	1 year
Min duration	1 hour	1 sec	1 sec	1 min
Max duration	1 year	15 years	6 days	5 hours
Avg. duration	67 days	1 year	40 mins	12 mins
Avg. duration [%]	18.4	7.22	0.03	0.002

overhead	BOOKS	WEBKIT	BTC	TAXIS
space	0.02%	0.04%	2.2%	0.09%
insertions	0.3%	0.4%	3.3%	3.1%
deletions	1.2%	0.2%	5.8%	3.1%

Overhead in space and maintenance costs for HINT

ACM SIGMOD/PODS International Conference on Management of Data, Berlin, Germany, June 22-27, 2025

Work supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "2nd Call for HFRI Research Projects to support Faculty Members & Researchers" (Project No. 2757)