MADALGO
CENTER FOR MASSIVE DATA ALGORITHMICS

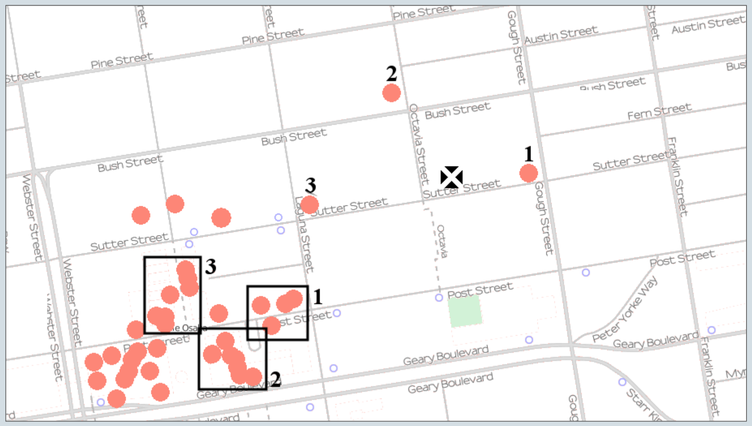

Anders Skovsgaard
Aarhus University

Danmarks Grundforskningsfond
Danish National Research Foundation

Finding Top-*k* Relevant Groups of Spatial Web Objects

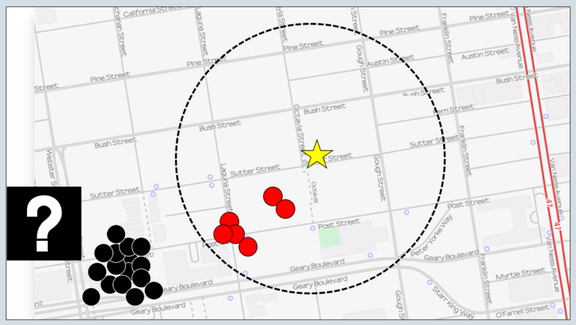
Motivation

- Top-*k* nearest single objects may not be relevant.
  - May be closed, too expensive or full.
- Instead return top-*k* most relevant groups of objects.
  - Gives the user a wide selection.



The Problem

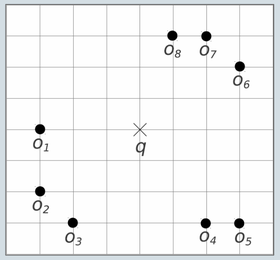
- **Group relevance:**
  - distance from the query location to the group.
  - diameter of the group.
  - number of objects in the group.
  - textual relevance between group keywords and objects in the group.
- **Candidate groups are all subsets of the dataset:**
  - hard to prune subtrees with exisisting work.

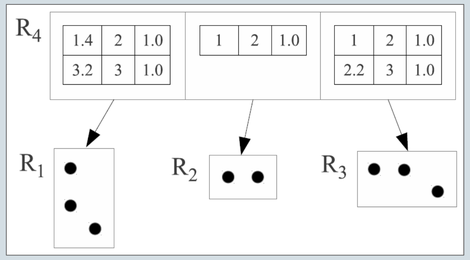


Group Extended R-Tree

The index:

Vocabulary: This is a mapping of each term to a GER-tree.
Trees: Objects are indexed in the GER-tree.
Compressed Histograms: Each entry in a non-leaf node in the GER-tree contains a compressed histogram, representing the child subtree.





Compressed Histograms

Conservative description of the subtree:
$\forall O \subseteq S\ (|O| > n_d \Rightarrow \textit{diameter}(O) > \textit{minDist})$

Full representation of the objects:

$minDist$	$n_d = \lvert O \rvert$	TR_{entry}	avg_{tf}
1.0	$\lvert\{o_7, o_8\}\rvert$	0.17	1.0
2.24	$\lvert\{o_7, o_6, o_8\}\rvert$	0.08	1.0
5.83	$\lvert\{o_2, o_4, o_3, o_1\}\rvert$	0.05	1.0
6.32	$\lvert\{o_7, o_6, o_4, o_8, o_1\}\rvert$	0.03	1.0
6.71	$\lvert\{o_2, o_4, o_3, o_8, o_1, o_5\}\rvert$	0.02	1.0
7.21	$\lvert\{o_7, o_2, o_6, o_4, o_3, o_8, o_1, o_5\}\rvert$	0.01	1.0

Compressed histogram:

$minDist$	n_d	avg_{tf}
1.0	2	1.0
2.24	3	1.0
5.83	4	1.0
6.32	8	1.0

Increasing the number of entries in the compressed histogram improves pruning.

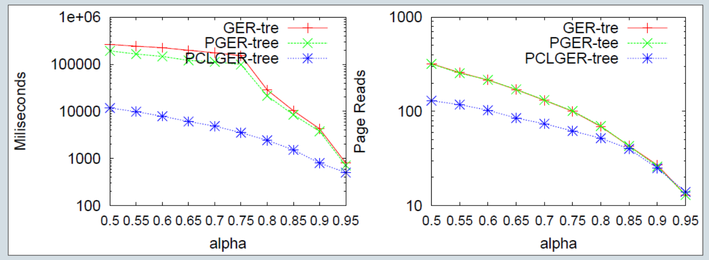
Query Processing

- **Groups may span across several entries:**
  - Conservative combination of entries.

Entries									Combined		
R_1			R_2			R_3					
3.2	3	1.0	1	2	1.0				4.0	5	1.0
3.2	3	1.0				2.2	3	1.0	3.6	6	1.0
			1	2	1.0	2.2	3	1.0	5.0	5	1.0
3.2	3	1.0	1	2	1.0	2.2	3	1.0	5.0	8	1.0

- Calculate *best case cost* based on the compressed histogram entries.
- At leaf level the *actual cost* is calculated.
- Terminate when actual cost is lower than any best case cost.

Experimental Results



α:	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.90	0.95
$\lvert G \rvert$:	9.2	8.7	7.7	7.0	6.3	5.5	4.4	3.5	2.2	1.2