Searchlight: Enabling Integrated Search and Exploration over Large Multidimensional Data

Alexander Kalinin  
Ugur Cetintemel  
Stan Zdonik
Brown University

Interactive Data Exploration (IDE)
Exploratory analysis: ad-hoc & repetitive
- Questions are not well defined
- “Interesting” can be complex

Human-in-the-loop operation
- Online answering

Constraint Programming Approach
Decision variables describe the object of search
- Left-most corner, side lengths

Constraints describe interesting properties
- Area 10,000 mi²
- Brightness more than 0.8

Two Sides of Data Exploration
Search complexity
- Search space is large
  Enumeration isn’t feasible
- Constraints are elaborate
  More than just ranges

Data complexity
- Large data sets (“big data”)
  Hard to fit in memory
- Expensive computations
  Functions over a lot of objects

Dynamic Solve-Validate Approach
CP Solver
- Searches in memory
- Works on a synopsis
- Extensive pruning

Validator
- Verifies candidates
- Accesses the disk
- Separate threads

Dynamic Balancing
- Solvers exchange sub-trees
- Validators request chunks
- Validators forward candidates

Table: Some Searchlight Results
<table>
<thead>
<tr>
<th>Approach</th>
<th>First result, s</th>
<th>Subsequent delays, s</th>
<th>Total time, s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Searchlight</td>
<td>4.8</td>
<td></td>
<td>5.13</td>
</tr>
<tr>
<td>CP</td>
<td>NA</td>
<td>NA</td>
<td>91</td>
</tr>
<tr>
<td>SciDB</td>
<td>301.3</td>
<td></td>
<td>945.3</td>
</tr>
</tbody>
</table>

8 node cluster 120 GB data 2GB memory

Contact: Alexander Kalinin, alexander_kalinin@brown.edu

Contact Information:
Intel  
MIT  
BROWN  
Carnegie Mellon University  
W  
UR  
Portland State University  
ISTC

SciDB Instance
Synopsis Array
Disk-resident Array
CP Solver
Candidate Solutions
Validator
Router
Where is Waldo?