Comparing Synopsis Techniques for Approximate Spatial Data Analysis

A. B. Siddique, Ahmed Eldawy, Vagelis Hristidis
University of California, Riverside
Spatial Data is Ubiquitous…

• The amount of spatial data is exponentially increasing

- 150 GB Weekly
- 50GB daily
- 60-80% Geo-referenced

Brain Simulation  ↑Crop Production  Event Detection  Geo- Advertisement
Spatial Data Synopsis

• Fast response algorithms are needed

• Data Synopsis:
  • Fast approximate results
  • Speed and accuracy trade-off

• Many spatial data synopsis

• Many common spatial operations

• Given a problem-setting, when to use which?
  • No readily-available answer
A Unified Environment

Big Dataset
In DFS

Memory Budget

Data Synopsis Generation

In-memory Data Synopsis

Synopsis-based Algorithms

Approximate Result

Evaluation of The Result

Full data Algorithms

Exact Result
Outline

Big Dataset
In DFS

Data Synopsis Generation

Synopsis-based Algorithms

Evaluation of The Result

Memory Budget

In-memory Data Synopsis

Approximate Result

Full data Algorithms

Exact Result
Synopsis Generation – Random Sample

• **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
• **Output:** Synopsis of the required size
Synopsis Generation – Stratified Sample

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size
Synopsis Generation – Uniform Histogram

- **Inputs:** Big dataset, Memory Budget, Synopsis algorithm
- **Output:** Synopsis of the required size

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>64</th>
<th>51</th>
<th>121</th>
<th>130</th>
<th>65</th>
<th>12</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>58</td>
<td>46</td>
<td>74</td>
<td>184</td>
<td>287</td>
<td>355</td>
<td>301</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>16</td>
<td>44</td>
<td>192</td>
<td>268</td>
<td>374</td>
<td>130</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>41</td>
<td>46</td>
<td>17</td>
<td>0</td>
</tr>
</tbody>
</table>
Synopsis Generation – Non-Uniform Histogram

- **Inputs**: Big dataset, Memory Budget, Synopsis algorithm
- **Output**: Synopsis of the required size

<table>
<thead>
<tr>
<th>121</th>
<th>49</th>
<th>34</th>
<th>55</th>
<th>49</th>
<th>77</th>
<th>186</th>
</tr>
</thead>
<tbody>
<tr>
<td>179</td>
<td>157</td>
<td>140</td>
<td>17</td>
<td>174</td>
<td>159</td>
<td>115</td>
</tr>
<tr>
<td>137</td>
<td>152</td>
<td>164</td>
<td>23</td>
<td>7</td>
<td>194</td>
<td>248</td>
</tr>
</tbody>
</table>


Outline

- Big Dataset
- In DFS
- Memory Budget
- Data Synopsis Generation
- In-memory Data Synopsis
- Synopsis-based Algorithms
- Approximate Result
- Full data Algorithms
- Exact Result
- Evaluation of The Result
Spatial Operations

• Spatial Operations studied in this work

Selectivity Estimation

K-Means Clustering

Spatial Partitioning
Selectivity Estimation (SE)

- Estimates the number of records in a given query range
- **Inputs:** Synopsis, Range Query
- **Output:** Number of records
SE – Based on Sample

• **Inputs:** Synopsis, Range Query
• **Output:** Number of records
### SE – Based on Histogram

- **Inputs:** Synopsis, Range Query
- **Output:** Number of records

#### Horizontal Prefix-sum

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>127</td>
<td>178</td>
<td>299</td>
<td>429</td>
<td>494</td>
<td>506</td>
</tr>
<tr>
<td>121</td>
<td>231</td>
<td>356</td>
<td>661</td>
<td>1078</td>
<td>1498</td>
<td>1811</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>427</td>
<td>924</td>
<td>1609</td>
<td>2403</td>
<td>2846</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>429</td>
<td>991</td>
<td>1717</td>
<td>2557</td>
<td>3017</td>
</tr>
</tbody>
</table>

#### Vertical Prefix-sum

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>127</td>
<td>178</td>
<td>299</td>
<td>429</td>
<td>494</td>
<td>506</td>
<td>545</td>
</tr>
<tr>
<td>121</td>
<td>231</td>
<td>356</td>
<td>661</td>
<td>1078</td>
<td>1498</td>
<td>1811</td>
<td>1899</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>427</td>
<td>924</td>
<td>1609</td>
<td>2403</td>
<td>2846</td>
<td>2934</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>429</td>
<td>991</td>
<td>1717</td>
<td>2557</td>
<td>3017</td>
<td>3105</td>
</tr>
</tbody>
</table>

**Results:**

- **+2403**
- **-494**
- **-427**
- **+178**

**Total:** 1606
SE – Based on Histogram

- **Inputs:** Synopsis, Range Query
- **Output:** Number of records

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>127</th>
<th>178</th>
<th>299</th>
<th>429</th>
<th>494</th>
<th>506</th>
<th>545</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>231</td>
<td>356</td>
<td>661</td>
<td>1078</td>
<td>1498</td>
<td></td>
<td>1811</td>
<td>1899</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>427</td>
<td>924</td>
<td>1609</td>
<td>2403</td>
<td></td>
<td>2846</td>
<td>2934</td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>439</td>
<td>991</td>
<td>1717</td>
<td>2557</td>
<td></td>
<td>3017</td>
<td>3105</td>
</tr>
</tbody>
</table>
K-Means Clustering (KC)

• An unsupervised learning problem, which tries to group objects having some kind of similarity into one cluster

• **Inputs:** Synopsis, Number of clusters

• **Output:** K-cluster centroids
KC – Based on Sample

- **Inputs:** Synopsis, Number of clusters
- **Output:** K-cluster centroids

\[ d(x, C)^2 \]
KC – Based on Histogram

- **Inputs:** Synopsis, Number of clusters
- **Output:** K-cluster centroids

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>63</td>
<td>64</td>
<td>51</td>
<td>121</td>
<td>130</td>
<td>65</td>
<td>12</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>46</td>
<td>21</td>
<td>184</td>
<td>267</td>
<td>355</td>
<td>301</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>44</td>
<td>192</td>
<td>268</td>
<td>374</td>
<td>130</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>16</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ f \times d(x, C)^2 \]
Full Dataset Algorithms

• State-of-the-art parallel algorithms for every operation

• Selectivity Estimation
  • Apache Spark’s Filter and count

• K-Means Clustering
  • Apache Spark MLlib’s Scalable K-Means++
Outline

Big Dataset
- In DFS

Memory Budget

Data Synopsis Generation
- In-memory Data Synopsis

Synopsis-based Algorithms
- Approximate Result

Evaluation of The Result

Full data Algorithms
- Exact Result

Exact Result

Memory Budget
## Experimental Setup

### Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Records</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>all-nodes</td>
<td>96 GB</td>
<td>2.7 billion</td>
<td>Points</td>
</tr>
<tr>
<td>edges</td>
<td>23 GB</td>
<td>70 million</td>
<td>Polygons</td>
</tr>
<tr>
<td>all-objects</td>
<td>92 GB</td>
<td>263 million</td>
<td>Mixed</td>
</tr>
<tr>
<td>Synthetic</td>
<td>51 GB</td>
<td>250 million</td>
<td>Rectangles</td>
</tr>
</tbody>
</table>

### Performance Metric

<table>
<thead>
<tr>
<th>Problem</th>
<th>Quality</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE</td>
<td>Absolute Relative Accuracy</td>
<td>Query Response Time</td>
</tr>
<tr>
<td>KC</td>
<td>Sum of Squared Error</td>
<td>Clustering Time</td>
</tr>
</tbody>
</table>
Selectivity Estimation: Quality Measure

Absolute Relative Accuracy

Selectivity Ratio = $10^{-3}$

Selectivity Ratio = $10^{-1}$

Accuracy

Memory Budget (Mb)

Memory Budget (Mb)

Edges dataset
Selectivity Estimation: Performance Measure

Query Response Time

Memory Budget = 20Mb

Memory Budget = 80Mb

Running time (ms)

Synthetic dataset
Full Dataset >55 seconds
K-Means Clustering: Quality Measure

Sum of Squared Error

Memory Budget = 2.16Mb

Memory Budget = 21.6Mb

All-nodes dataset
K-Means Clustering: Performance Measure

Clustering Time

All-nodes dataset
Guidelines: Selectivity Estimation

Dominant Objective

Exact Answer

Full Dataset

Approximate Answer

Query Size

Not Too Small

Non-Uniform Histogram

Too Small

Stratified Sample
Guidelines: K-Means Clustering

- **No. of Clusters**: Not too Large
  - Very Large: Full Dataset
  - Clustering time: Random Sample

- **Dominant Objective**: Clustering Quality
  - Uniform Histogram

- **Not too Large**: Objective
  - Not too Large: Objective
  - Clustering time: Random Sample

- **Very Large**: Data Set
  - Very Large: Data Set
  - Clustering time: Random Sample
Summary

Data Synopsis Generation

Synopsis-based Algorithms

Evaluation of The Result

Big Dataset

In-memory Data Synopsis

Memory Budget

Full data Algorithms

Exact Result

Approximate Result

Horizontal Prefix-sum

Vertical Prefix-sum

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>127</th>
<th>178</th>
<th>299</th>
<th>429</th>
<th>494</th>
<th>506</th>
<th>545</th>
</tr>
</thead>
<tbody>
<tr>
<td>121</td>
<td>231</td>
<td>356</td>
<td>661</td>
<td>1078</td>
<td>1498</td>
<td>1811</td>
<td>1899</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>427</td>
<td>924</td>
<td>1609</td>
<td>2403</td>
<td>2846</td>
<td>2934</td>
<td></td>
</tr>
<tr>
<td>132</td>
<td>258</td>
<td>429</td>
<td>991</td>
<td>1717</td>
<td>2557</td>
<td>3017</td>
<td>3105</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>64</th>
<th>51</th>
<th>121</th>
<th>130</th>
<th>65</th>
<th>12</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>46</td>
<td>74</td>
<td>184</td>
<td>287</td>
<td>355</td>
<td>301</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>44</td>
<td>192</td>
<td>268</td>
<td>374</td>
<td>130</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>41</td>
<td>46</td>
<td>17</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Dominant Objective

Approximate Answer

Query Size

Not Too Small

Non-Uniform Histogram

Exact Answer

Full Dataset

Stratified Sample

No. of Clusters

Not too Large

Uniform Histogram

Very Large

Full Dataset

Clustering Time

Random Sample

Cluster Quality

28
Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under grants IIS-1838222, IIS-1619463, and IIS-1447826.
Synopses Performance

Running time: edges

Running time: all-nodes

Memory Budget (Mb)
SPATIAL PARTITIONING (SP)

- Partitions into smaller-sized subsets
  - Balancing the sizes of the partitions
  - Maintaining the spatial locality
- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs
SP - Based on Sample

- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs
SP - Based on Histogram

- **Inputs:** Synopsis, Number of partitions
- **Output:** A set of MBRs

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>64</th>
<th>51</th>
<th>121</th>
<th>130</th>
<th>65</th>
<th>12</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>46</td>
<td>74</td>
<td>184</td>
<td>287</td>
<td>355</td>
<td>301</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>44</td>
<td>192</td>
<td>287</td>
<td>355</td>
<td>130</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>41</td>
<td>46</td>
<td>17</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
**SP - Based on Histogram Partial cells**

- **Inputs**: Synopsis, Number of partitions
- **Output**: A set of MBRs

<table>
<thead>
<tr>
<th></th>
<th>63</th>
<th>64</th>
<th>51</th>
<th>121</th>
<th>130</th>
<th>65</th>
<th>12</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>58</td>
<td>46</td>
<td>74</td>
<td>184</td>
<td></td>
<td>287</td>
<td>355</td>
<td>301</td>
<td>49</td>
</tr>
<tr>
<td>11</td>
<td>16</td>
<td>44</td>
<td>192</td>
<td></td>
<td>268</td>
<td>374</td>
<td>130</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>2</td>
<td>65</td>
<td>41</td>
<td>46</td>
<td>17</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>
SP: Quality Measures

Q3 for all-nodes

Q4 for edges

Q5 for all-objects
SP: Performance Measures

(a) $SP-F = 2275s$ for edges  
(b) $SP-F = 2993s$ for all-nodes
Summary of Results and Guidelines: SP

SP-NHP  \[\text{Optimize } Q1 \text{ and } Q5\]  Dominant Objective  Partitioning time  SP-UHP

Balance of all Qs

SP-RS(R*)