MIGIS: A HIGH PERFORMANCE QUERY SYSTEM FOR PATHOLOGY IMAGING ANALYTICS

Ablimit Aji\textsuperscript{1}, Qiaoling Liu\textsuperscript{1}, Fusheng Wang\textsuperscript{2,3} Ph.D., Tahisin Kurç\textsuperscript{2,3} Ph.D., Joel Saltz\textsuperscript{2,3} M.D., Ph.D.

\textsuperscript{1}Department of Mathematics & Computer Science
\textsuperscript{2}Department of Biomedical Informatics
\textsuperscript{3}Center for Comprehensive Informatics

Oct 10, 2012
Integrative Multi-Scale Biomedical Informatics

- Reproducible anatomic/functional characterization at gross level (Radiology) and fine level (Pathology).
- Integration of anatomic/functional characterization with multiple types of “omic”s information.
- Create categories of jointly classified data to describe pathophysiology, predict prognosis and response to treatment.

Emory In Silico Brain Tumor Research Center
Distinguishing Characteristics in Brain Tumor

Nuclear Qualities
- Round shaped with smooth regular texture
- Elongated with rough, irregular texture

Oligodendrogioma    Astrocytoma

Clinical data
Molecular data

Table:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>72.3</td>
</tr>
<tr>
<td>Perimeter</td>
<td>24.5</td>
</tr>
<tr>
<td>Circularity</td>
<td>0.26</td>
</tr>
<tr>
<td>Median</td>
<td>230</td>
</tr>
<tr>
<td>Variance</td>
<td>17.8</td>
</tr>
</tbody>
</table>

1 ... 4 5 ... 10
Oligo       Astro

Segmentation  Feature Extraction  Classification
Systematic Image Algorithm Evaluation

- High quality image analysis algorithms are essential to support biomedical research and diagnosis.
- Validate algorithms with human annotations
- Compare and consolidate different algorithm results
- Sensitivity study on algorithms’ parameters

Example: What are the distances and overlap ratios between markup boundaries from two algorithms?

Green: algorithm one
Red: algorithm two
• Quntum dot IHC allows for quantifying biomarker labeled signals and measuring multiple, stable biomarker signals simultaneously.

• We use QD IHC to map the activity of signal transduction pathways and transcriptional networks relative to the tumor microenvironment.

   *e.g.* for each stem cell, find the nearest blood vessel, compute the variation of intensity of each biological property associated with the cell in respect to the distance, and return the density distribution of blood vessels around each cell.
Spatial Centric Queries

- POINT
- CONTAINMENT
- WINDOW
- DENSITY
- SPATIAL JOIN
- NEIGHBOR
Both Data- and Compute-Intensive

• Explosion of derived data
  – $10^5 \times 10^5$ pixels per image
  – 1 million objects per image
  – Hundreds to thousands of images per study
  – **Big data demanding for high throughput**

• High computational complexity
  – Spatial queries include spatial refinement and spatial measurements, based on heavy duty geometric computation algorithms
  – **Demanding high performance**
High Performance Queries with MapReduce

• MapReduce is a parallel computing framework widely used for large-scale data analysis and queries
  – Invented by Google, open source version: Hadoop
  – Very easy to develop scalable applications
  – Two steps: *map* to partition problem into subproblems; *reduce* to combine results
  – Parallelization automatically managed by Hadoop

• Our approach:
  – Build efficient spatial query engine that can run easily deployed on clusters
  – Take advantage of MapReduce to run queries
MapReduce Based Spatial Cross-Matching Query Processing

- **Staging:**
  - result files redistributed across HDFS as blocks.
  - small tiles merged and metadata added into records.

- **Map:**
  - identify records of same tiles to form tasks.

- **Reduce**
  - execute queries with real-time spatial query engine.
  - aggregate query results.
Real-Time Spatial Query Engine (RESQUE)

- Index building on demand (low overhead).
- Query pipelines to combine multiple steps of query processing.
- Support of spatial join, multi-way spatial join, nearest neighbor, and highest density queries, and extensible for new ones.
- Able to run in parallel with decoupled spatial processing in a distributed computing environment.

Example: Two-way Spatial Join

Boundary File1 → Bulk R*-Tree Building → R*-Tree File1 → Spatial Join Algorithm → Geometry Refinement → Spatial Measure → Result File

Boundary File2 → Bulk R*-Tree Building → R*-Tree File2

Real-time spatial querying engine (RESQUE)
Nearest Neighbor Query Processing Workflow
MIGIS: Medical Imaging GIS

Image Analysis

Analytical Results

Query System

Spatial and Pattern Discovery Language (SPDL)

YSMART-S
- SPDL to MR Translator
- MR Job Optimizer
- Cached Queries

Partitioner
- Spatial Index Builder

Aggregator
- Spatial Query Processor

HDFS

Hadoop
Real-Time Spatial Query Engine: Performance

- **R*-Tree building**: ~16%; **R*-Tree join**: ~84%
- **Storage**: compression of boundaries in R*-Tree leave nodes as chain code: save 42% space
- **Scalability**: multiple small spatial index has same performance as combined big index.
System Scalability

- on average 15 sec/image (join query)
- 22 min on PostGIS
Data Skew & Query Optimization

- Data skew is common.
- Hash partitioning doesn’t promise even distribution.
- Cost based query optimization.
Effective management of large scale data is one of the pressing challenges for next generation integrative biomedical study.

We provide a high performance querying system for pathology analytical imaging.

The system can efficiently support various statistical, spatial queries and complex analytical queries.

The system is cost effective, scalable and easy to use.
Thank you!

MIGIS: A High performance Query system for Pathology Imaging analytics

Ablimit Aji, Qiaoling Liu, Fusheng Wang, Tahisin Kurç, Joel Saltz