New Trends in Database Systems

Ahmed Eldawy
Spatial and Spatio-temporal data
What is spatial data

Geographical data

Medical images

Astronomical data

Trajectories
Application of spatial data

- Tracking of infectious disease
- Geo-targeted advertising
- Geographic Information Systems (GIS)
- Identifying local events/groups
- Disaster recovery/eviction plans
- Routing
Query processing

- Point selection
- Range selection
- Nearest neighbor
- Spatial join
Indexing

▷ A better organization of records to speed up the query processing

Why do we need new indexes?

R-tree

Quad-tree
Selectivity estimation

- Estimates the size of the answer without having to run the query
- Useful for query optimization
- e.g., range selection queries

💡 Cost model for spatial join on big data
Road networks

- Road intersections as graph vertices
- Road segments as (weighted) edges
- Planar graphs
- Adds topology semantics to queries
- e.g., nearest neighbor and clustering
-💡 Build and visualize a 3D road network
Big Spatial Data

- Four V’s of big data
- Volume
  - Partitioning and distributed query processing
- Velocity
  - In-memory indexes
  - Flushing policy
- Variety
  - Combine vector and raster data
- Veracity
  - Inherent errors in location data
Project ideas for BSD

- A benchmark for BSD systems
  - Real data is available
  - A mix of range and spatial join queries
  - Run on a few systems
- Implement a new query on a BSD system
  - spatial join in AsterixDB
  - KNN join on SpatialHadoop
- Aggregated visualization of BSD
  - e.g., color code cities by number of tweets
  - Integrate AsterixDB with a visualization server
Volunteered Geographic Information (VGI)

- Collection of geographic data by volunteers
- OpenStreetMap

- How to take user locations into account?
- Identifying experts
Ridesharing

- Uber

Matching drivers and passengers

Tracking vehicles

Figure from http://www.eia.gov/todayinenergy/detail.php?id=13531
Spatial keyword queries

- Keywords are assigned to each record

- Expand queries to incorporate keywords
  - Range queries: Find all *restaurants* for *kids* within 10 miles
  - Nearest neighbor queries: Find the nearest *gas station* that accepts *credit card*
  - Join queries: Find *excellent* *schools* and *2br* *houses* that are within a distance of five miles
Indoor Environment

- Motivated by new technologies
  - WiFi and Bluetooth localization
  - RFID
  - NFC

- Challenges
  - 3D localization
  - Indoor maps
  - Indoor routing
Geo-image browser
Geo-image browser
Geo-image browser
Emerging Applications
Traditional Applications

› Relational data schema
  › Tables, columns, and rows
  › Normalization

› Relational operations
  › Select, project, join, group and aggregate

› Descriptive query language
  › SQL

› ACID transaction guarantees
New Requirements

- Schema-less
- Hierarchical formats
- New operations
  - Closures
  - Similarities
- Scientific language
  - R, Matlab
- Weak or no transaction guarantees
Social Networks

- Eventual consistency
- High volume + High velocity
Graph processing

- Social networks
- Web graphs
- Call Detail Records (CDR)
- Road networks
- Brain simulation data
- Knowledge base and RDF data

💡 A benchmark for big graph systems, e.g., GraphX, Giraph, and Pregelix
Data cleaning

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
<th>Affiliation</th>
<th>ZIP code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark</td>
<td><a href="mailto:mark@ucr.edu">mark@ucr.edu</a></td>
<td>UC Riverside</td>
<td>922521</td>
</tr>
<tr>
<td>Anthony</td>
<td><a href="mailto:james@uci.edu">james@uci.edu</a></td>
<td>UC Riverside</td>
<td>92697</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Area code</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark</td>
<td>UCR</td>
<td>951</td>
<td>555-5555</td>
</tr>
<tr>
<td>Tony</td>
<td>UC Irvine</td>
<td>949</td>
<td>555-5555</td>
</tr>
</tbody>
</table>

- How many errors can you spot?
- Can you fix any of them?
-💡 Data cleaning of big data on Spark
Crowd sourcing

- Outsource *hard* problems to people
- e.g.,
  - Select all pictures with a waterfall
  - Do these two pictures show the same person?
  - Which query result is more relevant?

- Challenges
  - Minimize cost
  - Achieve a high accuracy
  - Task allocation
Visualization

➢ Explore big data through visualization

➢ Challenge: Make these visualizations scalable and interactive
Challenge: Keep data safe while providing scalable computing
Summary

- Spatial and spatio-temporal data
  - Query processing
  - Big spatial data
  - Indexing
  - Road networks
- Emerging applications
  - Social networks
  - Graph processing
  - Visualization
  - Data cleaning
Thank You!