Provenance for Interactive Visualizations

Fotis Psallidas
Columbia University
fotis@cs.columbia.edu

Eugene Wu
Columbia University
ew2493@columbia.edu

Background

- Data scientists need to explore large volumes of information to gain data insights as fast as possible.
- Interactive visualizations bring powerful ad-hoc and focused analysis to both technical and non-technical data scientists.
- Interactive visualizations are increasingly adopted across domains.
- Machine learning toolkits (e.g., Tensorflow and R Studio).
- Decision support systems (e.g., Endeca, Tableau, and PowerPivot).
- Knowledge exploration (e.g., TimeMachine and Yago Explorer).
- News (e.g., Interactive Stories at New York Times).

Problem Statement

Specify and optimize interactive visualizations:

- Previous approaches using data visualization systems or manual implementation using popular toolkits that scale with the Web.
- Data visualization systems: limited interactions yet easy to use.
- Tableau: interactions for OLAP queries.
- Crossfilter: linked selection interactions for correlated dimensions.
- Manual implementations: hard to optimize, maintain, extend, and reason about.

The DVMS Project

- A declarative, relational approach to interactive visualizations.
- Main win: push down optimization and design choices into the database system with readily available guarantees and optimizations.
- DVMS ecosystem: several extensions to the database box for the optimization and design of interactive visualizations.
- Provenance subsystem for the specification and optimization of interactive visualizations.
- Other extensions:
  - Asynchronous interactions using concurrency control ideas.
  - Streaming for the optimization of near-interactive visualizations.
  - Precision interfaces for recommendation of new designs.
  - and many more…

Static Visualizations

- Data from a database converted to visual marks using SQL queries (i.e., as database views).
- Visual marks (encoded as views) are rendered using mark-specific render functions.
- Provenance subsystem: database system with readily available guarantees and optimizations.

Interactive Visualizations using Provenance

- Selection of marks: (user event stream → visual marks).
- Linked brushing: Backward trace selection to base data and refresh the coordinated view based on the traced subset.

Optimizing Interactions through Provenance

- Instead of going all the way back and forward, keep mappings (provenance materialization).
- Main problem:
  - Materializing and indexing provenance can be expensive.
  - Many possible options for materialization. Which one to pick and when?

Evaluation and Results

- Specification:
  - Completeness on several interaction taxonomies.
- Support for many novel techniques due to provenance availability.
- Interactive debugging of applications.
- Deconstruction and restyling of visualizations.
- Data explanation.
- Provenance visualization.
- Optimization:
  - Preliminary results: 20-150% improvements on provenance materialization latency against state-of-the-art provenance subsystems.
  - Expecting improvements on interaction latency with novel materialization options (i.e., data skipping and delta propagation).

DVMS: Research Prototype

DeVIL Program

- Static Analysis
- Offline Optimizer
- Interaction Manager
- Executor
- Optimizer
- Storage Manager

Evaluation and Results

DVMS: Research Prototype