Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment

Zaina Shaik, Role: Paper Author

Problem Area Background

- Data sets can have missing / extreme / duplicate / inaccurate values
- Can harm the analysis of data
- Costs billions of dollars annually to fix inaccurate analyses
- Identifying data errors requires human judgement
- Difficult to contextualize the anomaly and choose how to best move forward with visualizations (subset of data, type of visualization, sorting)

Introducing Profiler

- RQ: How do you create a system to find and support data anomalies?
- Profiler: visual analysis tool used for identifying and assessing data quality issues
- Finds potential data anomalies with type inference and data mining routines
- Suggests multi-view visualizations to assist data analysts contextualize data
- Extensible system architecture:
  - Modular architecture with statistical and visual analysis
- Automatic view suggestion:
  - Uses mutual information to give recommendations
- Scalable summary visualizations
  - Includes brushing and linked selections

Related Work - Classifying Data Anomalies

- Previous work focused on errors that require some human intervention to assess
- Profiler focuses on:
  - Missing data
  - Erroneous data
  - Inconsistent data
  - Extreme values
  - Key violations

Related Work - Data Cleaning Tools / Visual Analysis

- Previous work focused on data integration or entity resolution
- Profiler looks at data anomalies in a single table and can detest a broader range of discrepancies
- Ex: Google Refine
- Integrated with Wrangler

- Previous work has looked at creating multi-dimensional views
- Profiler automatically recommends subsets and summary visualizations
- Ex: Tableau

User Interface

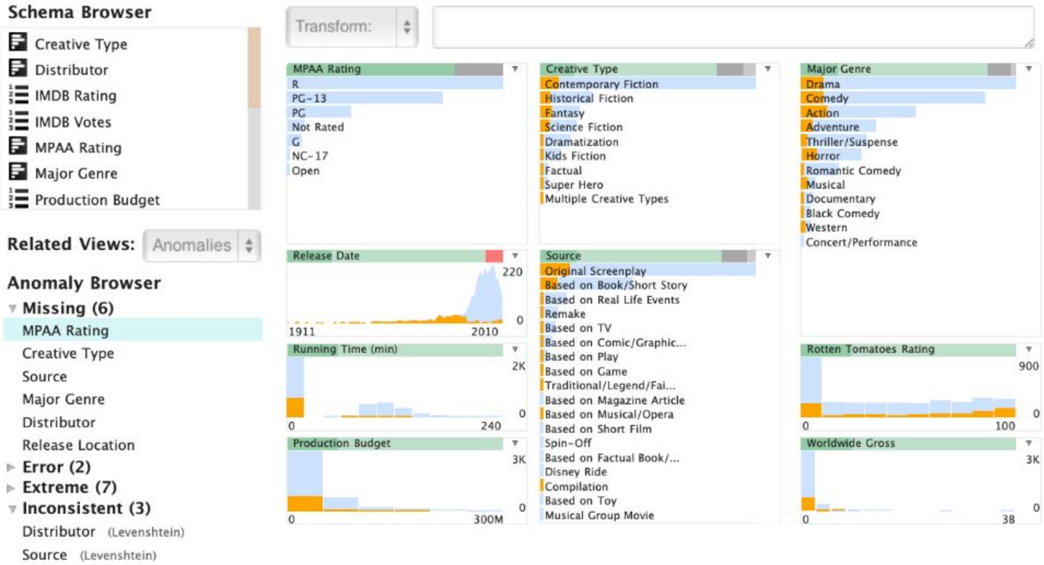


Figure 1: The Profiler User Interface. The UI contains (clockwise from top-left): (a) schema browser, (b) formula editor, (c) canvas of linked summary visualizations, and (d) anomaly browser. Profiler generates a set of linked views for each identified anomaly. Here, we investigate possible causes of missing MPAA movie ratings. The grey bar above the MPAA rating chart indicates missing values; we select it to highlight matching records. The Release Date chart shows that missing ratings correlate with earlier release dates.

User Interface

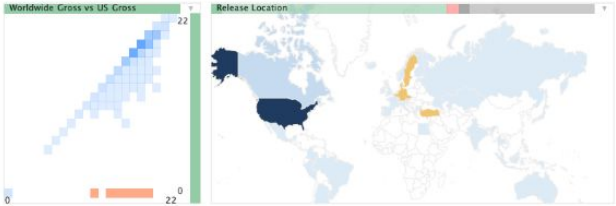


Figure 3: Map assessing 2D outliers in a binned scatter plot. Selected in the scatter plot are movies with high Worldwide Gross but low US Gross (in orange). Linked highlights on the map confirm that the movies were released outside of the US.

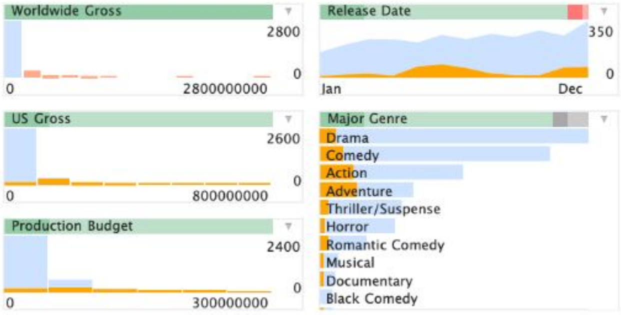
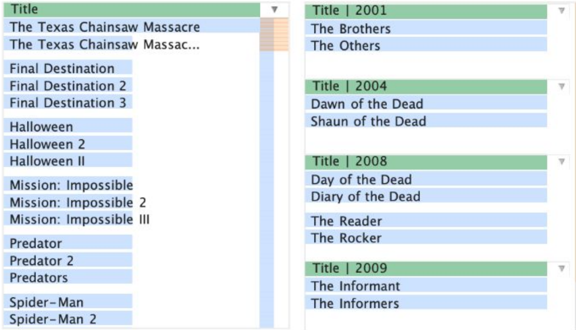


Figure 2: Automatically generated views to help assess Worldwide Gross. Worldwide Gross correlates with high US Gross and Production Budgets. High gross also coincides with Action & Adventure movies and the Summer & Winter seasons. Profiler chose to bin Release Date by month instead of by year.



Title

The Texas Chainsaw Massacre
The Texas Chainsaw Massac...

Final Destination
Final Destination 2
Final Destination 3

Halloween
Halloween 2
Halloween II

Mission: Impossible
Mission: Impossible 2
Mission: Impossible III

Predator
Predator 2
Predators

Spider-Man
Spider-Man 2

Title | 2001

The Brothers
The Others

Title | 2004

Dawn of the Dead
Shaun of the Dead

Title | 2008

Day of the Dead
Diary of the Dead

The Reader
The Rocker

Title | 2009

The Informant
The Informers

Figure 4: Conditioned duplicate detection. Left: Movie titles clustered by Levenshtein distance reveal over 200 potential duplicates. Right: Conditioning the clustering routine on 'Release Year' reduces the number of potential duplicates to 10.

System Architecture

- Data Tables
- Type Registry
- Detector
- Recommender
- View Manager

Data Tables

- Memory-resident column-oriented relational database
- Filtering, aggregation, generating derived columns
- Relaxed-type system
    - Flags values that deviate as inconsistent

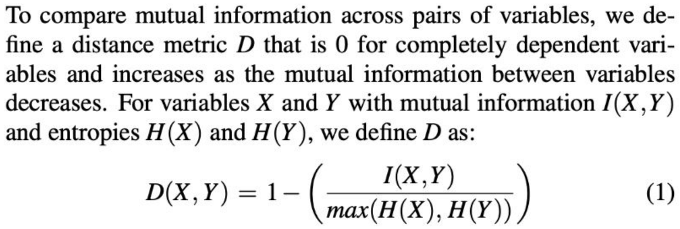
Type Registry

- Binary verification function
- Type definition
  - Type transforms
  - Group-by functions
- Type inference
  - Minimum Description Length Principle

Detector

- Applies collection of type-specific data mining routines to identify anomalies
- Derives features, extracts with generators
- Analyze with detection routines
  - Class column
  - Certainty column
- Organization
- Routines
  - Missing value detection
  - Type verification
  - Clustering
  - Univariate outlier detection
  - Frequency outlier detection

Recommender



To compare mutual information across pairs of variables, we define a distance metric D that is 0 for completely dependent variables and increases as the mutual information between variables decreases. For variables X and Y with mutual information $I(X,Y)$ and entropies $H(X)$ and $H(Y)$, we define D as:

$$D(X,Y) = 1 - \left(\frac{I(X,Y)}{max(H(X),H(Y))} \right) \quad (1)$$

- Utilized mutual information
- ViewToColumn
  - Function that converts a view specification into a column of group ids
- Anomaly-oriented views
  - Best predict the class column
- Value-oriented views
  - Predict the group ids generated by primary view specification

View Manager

- Summary Visualizations
    - Histograms
    - Area charts
    - Choropleth maps
    - Binned scatter plots
    - Bar charts
    - Grouped bar charts
    - Data quality bars
    - Uses small multiples
    - Table display
- Linked Highlighting

Evaluation

- Informal evaluations with water quality data, disasters database, obesity data, a quality-of-life inted, and public government data
- Looked at Profiler's results and observed
- Disaster database:
    - 11 columns, 13 data quality issues
    - For columns with missing values, recommends columns with co-occurrences of missing values
    - Took a few minutes
- World Water Monitoring Data
    - 34 columns, 35 quality issues
    - Flagged possible duplicates

Summary

- Profiler: extensible system for assessing data quality
- Wide range of data type support
- View recommendation model assesses data mining routines and suggests visual data summaries
- Linked summary helps evaluation
- Profiler can reduce time spent finding data quality issues, letting users make a deeper analysis

Future Work

- Conduct a more thorough evaluation with controlled studies and public web deployments
- Create a tool for defining custom types to include detectors and visualizations for additional data types
- Explore hybrid approaches with server-side aggregation and client-side interactive querying
- Methods that could account for conditional dependencies between sets of columns

Thank you!