Scorpion

Explaining Away Outliers in Aggregate Queries

eugene wu and sam madden
MIT
More than 6.1 million private sector jobs added in the past 35 months.
<table>
<thead>
<tr>
<th>Table</th>
<th>Split</th>
<th>Aggregate</th>
<th>Visualize</th>
</tr>
</thead>
</table>

Diagram showing the process from Table to Split, then Aggregate, and finally Visualization.
SELECT sum(cost) 
FROM expenses 
GROUP BY country
SELECT sum(cost) 
FROM expenses 
GROUP BY country
SELECT sum(cost)
FROM expenses
GROUP BY country
SELECT sum(cost)
FROM expenses
GROUP BY country
Given
Outlier and normal results

Understand Why

```
SELECT sum(cost) FROM expenses GROUP BY country
```
Given
Outlier and normal results

What input properties
caused the outliers?

*most* caused the outliers?

causd outliers but didn’t affect normal outputs?

```
SELECT sum(cost) 
FROM expenses 
GROUP BY country
```
Can’t Touch This
I DON'T ALWAYS SOLVE PROBLEMS

BUT WHEN I DO, I DATABASE
Provenance

SHOW ME THE Data!
Provenance

SELECT SUM(cost)
FROM sam’s bank account
SELECT SUM(cost)
FROM sam’s bank account
SELECT SUM(cost) 
FROM sam’s bank account
SELECT SUM(cost) 
FROM sam’s bank account

Darn!
Ya caught me
WHAT'S THE POINT

SOONER OR LATER WE ALL CROAK
SELECT SUM(cost) 
FROM sam’s bank account

Filter for 
“most influential”
Provenance

Faceting

Provenance

Faceting

Dimensionality :
Dealing with multiple outliers?

Scorpion!
Given
Outlier and normal results

Understand Why
Given
Outlier and normal results

Find
Predicates correlated with outliers

\[ \text{Desc} = \text{“toilets”} \]
Given
Outlier and normal results

Find
Predicates correlated with outliers

s.t.
Removing predicate from inputs “fixes” outliers & maintains normal results

Desc = “toilets”
Given
Outlier and normal results

Find
Predicates correlated with outliers

s.t.
Removing predicate from inputs “fixes” outliers & maintains normal results

Expenses

USA  China  Italy

Dese = “toilets”
Given
Outlier and normal results

Find
Predicates correlated with outliers

s.t.
Removing predicate from inputs “fixes” outliers & maintains normal results

Dese = “toilets”
Formalize “influence” as metric
Predicate search heuristics
Some results
Desc = “toilet”
$p(T)$
p(T)
$p(T)$
\[ \Delta \text{output} \]

\[ p(T) \quad \text{---} \quad |p(T)| \]
\[ \frac{\Delta \text{output}}{|p(T)|} \] Influence Metric
\[
\frac{\Delta f(x)}{\Delta x} \quad \text{Sensitivity Analysis}
\]

\[
\frac{\Delta \text{output}}{|p(T)|} \quad \text{Influence Metric}
\]
\[ \Delta \text{Output} \]

\[ \frac{\Delta \text{output}}{|p(T)|} \]

"High vs Low"

\[ |p(T)| \]

\[ \Delta \text{Normal} \]

Multiple Outputs
Δ Output

"High vs Low"

Δ output

|p(T)|

Δ Normal

Δ output • V

|p(T)|

Multiple Outputs
Δ Output

|p(T)|

Δ output

Δ output • V

|p(T)|

Δ output • V

|p(T)|^c

“High vs Low”

Δ Normal

Multiple Outputs
Δ Output

Δ Output

Δoutput

|p(T)|

Δ output • V

|p(T)|

Δ Normal

Δ outlier • V

|p(T)|

Δ Normal

Multiple Outputs
influence(p)
Formalize “influence” as metric

Predicate search heuristics

Some results
\[ p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p) \]
$$p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p)$$

$$O(\text{agg}(T-p(T)))$$
\[ \text{SUM} \{1,2,3,4,5\} = 15 \]

\[ p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{influence}(p) \]

\[ O(\text{agg}(T-p(T))) \]
\[ p^* = \underset{p \in \text{predicates}}{\arg \max} \text{influence}(p) \]

\[ \text{SUM}\{1,2,3,4,5\} = 15 \]
\[ \text{p} \]

\[
\text{SUM} \{1,2,3,4,5\} = 15
\]

\[
\{4,5\}
\]

\[ p^* = \arg \max_{p \in \text{predicates}} \text{influence}(p) \]

\[ O(\text{agg}(T-p(T))) \]
\[ p^* = \text{argmax}_{p \in \text{predicates}} \text{influence}(p) \]

\[ \text{SUM}({1,2,3,4,5}) = 15 - \{4,5\} \]

\[ \text{SUM}({1,2,3}) = 6 \]

\[ O(\text{agg}(T-p(T))) \]
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

\[ O(\text{exponential}) \quad \text{O(agg(T-p(T)))} \]
\[ p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p) \]

- \( O(\text{exponential}) \)
- \( O(\text{agg}(T - p(T))) \)
\[ p^* = \arg \max_{p \in \text{predicates}} \text{influence}(p) \]

- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)

**Incrementally removable**
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

\[ \sum\{1,2,3,4,5\} = 15 \]

\[ O(\text{exponential}) \]

\[ O(\text{agg}(p(T))) \]

Incrementally removable
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

\[ \text{SUM}\{1,2,3,4,5\} = 15 \]

\[ 15 - \text{SUM}\{4,5\} = 6 \]
\[ p^* = \text{argmax}_{p \in \text{predicates}} \text{influence}(p) \]

- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)

- Incrementally removable
\[ p^* = \operatorname{argmax}_{p \in \text{predicates}} \, \text{influence}(p) \]

\[ \mathcal{O}(\text{exponential}) \]

\[ \mathcal{O}(\text{agg}(p(T))) \]

Incrementally removable
Least influence  \[ \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \bullet \]  Most influence

\[ p^* = \operatorname{argmax}_{p \in \text{predicates}} \text{influence}(p) \]

- Independent
- Incrementally removable

\[ O(\text{exponential}) \]
\[ O(\text{agg}(p(T))) \]
\( p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \)

Least influence \[\cdots\] \[\cdots\] \[\cdots\] \[\cdots\] \[\cdots\] Most influence

\( O(\text{exponential}) \quad \text{Independent} \quad O(\text{agg}(p(T))) \quad \text{Incrementally removable} \)
Least influence \[ \cdots \] Most influence

\[
p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p)
\]

- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)

**Independent** Incrementally removable
\[ p^\ast = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p) \]

\[ O(\text{exponential}) \quad \text{Independent} \quad O(\text{agg}(p(T))) \quad \text{Incrementally removable} \]
\[
p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p)
\]

- Top Down
- Independent
- Incrementally removable

\[O(\text{exponential}) \quad O(\text{agg}(p(T)))\]
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)

- Top Down
- Independent
- Incrementally removable
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

- Top Down
- Independent
- Incrementally removable
\[ p^* = \underset{p \in \text{predicates}}{\text{argmax}} \text{ influence}(p) \]

- \[ O(\text{exponential}) \]
- \[ O(\text{agg}(p(T))) \]

**Top Down**  **Independent**  **Incrementally removable**
\[ p^* = \underset{p \in \text{predicates}}{\text{argmax}} \quad \text{influence}(p) \]

\[ O(\text{exponential}) \]

\[ O(\text{agg}(p(T))) \]

- Top Down
- Independent
- Anti-monotonic
- Incrementally removable
\[ p^* = \text{argmax}_{p \in \text{predicates}} \text{influence}(p) \]

- Top Down
- Independent
- Anti-monotonic

- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)

- Incrementally removable
\[ p' \subset p \]
\[ \text{influence}(p') \leq \text{influence}(p) \]

\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

- Top Down
- Independent
- Anti-monotonic

\[ O(\text{exponential}) \]
\[ O(\text{agg}(p(T))) \]

Incrementally removable
\[ p^* = \text{argmax}_{p \in \text{predicates}} \text{influence}(p) \]

- **Top Down**
- **Independent**
- **Incrementally removable**
- **Bottom Up**
- **Anti-monotonic**

Complexity:
- \( O(\text{exponential}) \)
- \( O(\text{agg}(p(T))) \)
\( p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \)

- **Top Down**
- **Independent**
- **Bottom Up**
- **Anti-monotonic**
- **Incrementally removable**

\( O(\text{exponential}) \)
\[ p^* = \arg\max_{p \in \text{predicates}} \text{influence}(p) \]

\[ O(\text{exponential}) \]

Top Down \hspace{2cm} Independent \hspace{2cm} Incrementally removable

Bottom Up \hspace{2cm} Anti-monotonic
Formalize "influence" as metric
Predicate search heuristics

Some results
SELECT \sum(Y) \text{ GROUP BY } X
SELECT \text{sum}(Y) \ \text{GROUP BY} \ X
SELECT \text{sum}(Y) \text{ GROUP BY } X
SELECT \( \text{sum}(Y) \) GROUP BY X
influence metric that is accessible to end-users for

Data cleaning

Data exploration

Provenance reduction
scorpion

eugenewu@mit.edu
scorpion

eugenewu@mit.edu
C-parameter

\[
\Delta \text{output} \cdot V \cdot \frac{1}{|p(T)|^c}
\]