A Probabilistic Approach to Data Summarization

Laurel Orr, Magdalena Balazinska, and Dan Suciu
DB Research Day 2015

“After careful consideration of all 437 charts, graphs, and metrics, I’ve decided to throw up my hands, hit the liquor store, and get snockered. Who’s with me?!”
Original Database
(All Flights in US)

Summary

What are the most popular flights?

Flights from Los Angeles to San Diego
Existing Summarization Techniques

Sampling

Random, Stratified, Weighted, ...

Aggregation

Online, Error Bounded (BlinkDB)
Flights (origin, destination, fl_time, ...) ~ 2.6 GB

**Sampling**

```
SELECT origin, COUNT(*)
FROM Flights
GROUP BY origin;
```

Full Query Time: 20 sec

```
SELECT *
FROM Flights
WHERE origin='SEATTLE, WA'
LIMIT 10;
```

Full Query Time: 0.4 sec

```
SELECT origin, COUNT(*)
FROM Flights
WHERE dest = 'LAUREL, MS'
AND fl_time < 120
GROUP BY origin;
```

Full Query Time: 30 sec

**Aggregation**

- ✔️
- ✔️
- X
- ?
IDEA
Find a compact, \textit{probabilistic} representation of our database

Flights with high probability of existence \quad = \quad \text{Popular}

By knowing the probability of relations and tuples, we can answer queries probabilistically
The Simplest Summary

Assume there is some concrete relation $R(A, B)$, and you summarized $R$ by its active domain and cardinality.

Given this summary alone, what are the possible relations $R$ could have been (possible worlds of $R$)?
Possible World Semantics

active domain

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

$n = 2$

slotted instance

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$X = 4$

$X = 4$

$X = 4$

$X = 4$

$X = 4$

= 16 Possible Instances

$$\sum_{I \in PWD} \Pr(I) = 1 \quad \Rightarrow \quad \Pr(I) = \frac{1}{16}$$

set of all possible instances (stand for Possible WorLDs)
Possible World Semantics

active domain

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_1$</td>
<td>$b_1$</td>
</tr>
<tr>
<td>$a_2$</td>
<td>$b_2$</td>
</tr>
</tbody>
</table>

n = 2

slotted instance

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>id1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>id2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

= 16 Possible Instances

$$\Pr((a_1, b_1)) = \sum_{I \in PWD} \frac{1}{16} = \frac{7}{16}$$
Adding Constraints

active domain

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>a₁</td>
<td></td>
<td>b₁</td>
</tr>
<tr>
<td>a₂</td>
<td></td>
<td>b₂</td>
</tr>
</tbody>
</table>

n = 100

| σ_{R.A = a₁(R)} | = 70
| σ_{R.A = a₂(R)} | = 30
| σ_{R.A = a₁ \land R.B = b₁(R)} | = 40

probabilistic instance

\[ E[|σ_{I.A = a₁(I)}|] = 70 \]
\[ E[|σ_{I.A = a₂(I)}|] = 30 \]
\[ E[|σ_{I.A = a₁ \land I.B = b₁(I)}|] = 40 \]

\[ \sum_{I \in PWD} |σ_{I.A = a₁(I)}| \Pr(I) = 70 \]
\[ \sum_{I \in PWD} |σ_{I.A = a₂(I)}| \Pr(I) = 30 \]
\[ \ldots \]
\[ \sum_{I \in PWD} |σ_{I.A = a₁ \land I.B = b₁(I)}| \Pr(I) = 40 \]

How can we solve for \( \Pr(I) \)?
Principle of Maximum Entropy

The Principle of Maximum Entropy states that subject to prior data, the probability distribution which best represents the state of knowledge is the one that has the largest entropy.

In other words, you want to maximize

\[- \sum_{I \in PW \in D} Pr(I) \ast \log(Pr(I))\]

over all possible worlds.
More Formally

$R(A_1, \ldots, A_m), |R| = n$

$D_i = \text{distinct domain of } A_i,$

$\text{Tup} = \{D_1 \times D_2 \times \ldots \times D_m\},$

$\Phi = \text{set of equality predicates } \phi$

\[
Pr(I) = P^{-n} \prod_{\phi \in \Phi} \alpha_{\phi}^{\sigma_{\phi}(R)}
\]

\[
P = \sum_{t \in \text{Tup}} \prod_{\phi \in \Phi | \phi(t) = \text{true}} \alpha_{\phi}
\]

all possible tuples in our active domain
To include constraints on each $\phi$

$$s_R(\phi) = |\sigma_\phi(R)| = E[|\sigma_\phi(I)|]$$

We can show

$$s_R(\phi) = \frac{n\alpha_\phi P_{\alpha_\phi}}{P}$$

derivative of $P$ with respect to $\alpha_\phi$

To solve, maximize the potential function by gradient descent

$$\Psi = \sum_{\phi \in \Phi} \ln(\alpha_\phi) s_R(\phi) - \ln(P^n)$$
Query Transformation

Aggregates: take expected value

SELECT origin, COUNT(*)
FROM Flights
GROUP BY origin;

GROUP BY + COUNT(*)

For each origin $o$

$E[|\sigma_{origin=o}(Flights)|]$

$E[|\sigma_{\phi}(I)|] = \frac{n\alpha_{\phi}P_{\alpha_{\phi}}}{P}$

An equation in terms of the $\alpha$’s we have calculated and stored
Optimizations

\[
P = \sum_{t \in \text{Tuple}} \prod_{\phi \in \Phi | \phi(t) = \text{true}} \alpha_{\phi}
\]

all possible tuples in our active domain

1. Factorize \( P \) (solve 1D predicates independently)
2. Add relevant 2+D predicates (ex: \([A = a1 \land B = b1]\))
3. Remove tuples that don’t exist

\[
P^* = P - \sum_{t \in (\text{Tuple} - R)} \prod_{\phi \in \Phi | \phi(t) = \text{true}} \alpha_{\phi}
\]

4. Change Basis (for correlations)
   new attribute \( AB = f(A, B) \)
   (ex: \( AB = A - B \))
Experiment with TPC-H

```
SELECT order_date, ship_date, COUNT(*)
FROM orders JOIN lineitem
GROUP BY order_date, ship_date;
```

\[
\text{Error} = \frac{|Est - True|}{Est + True}
\]
Change Basis: order_date – ship_date
Conclusion

• Introduced new way to summarize and approximately query massive datasets
  – Complements sampling and approximate aggregation
• Allows fine grained control over which attributes and values get summarized
• Encouraging preliminary results
• Still need to better address scalability and expand query language
• Need to understand how best to choose statistics