Data Provenance for Query Result Explanation

Yi Chen
Arizona State University

Joint work with Subbarao Kambhampati, Garrett Wolf, Hemal Khatri, Bhaumik Chokshi, Jianchun Fan, Ullas Nambiar
Challenges in Querying Web: Incomplete Data

- Incomplete Entry
- Extraction Inaccuracy
- Schema Heterogeneity

<table>
<thead>
<tr>
<th>Website</th>
<th># of attributes</th>
<th># of tuples</th>
<th>incomplete tuples</th>
<th>body style</th>
<th>engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>autotrader.com</td>
<td>13</td>
<td>25127</td>
<td>33.67%</td>
<td>3.6%</td>
<td>8.1%</td>
</tr>
<tr>
<td>carsdirect.com</td>
<td>14</td>
<td>32564</td>
<td>98.74%</td>
<td>55.7%</td>
<td>55.8%</td>
</tr>
</tbody>
</table>
Wanted a ‘sedan’ priced around $7000

A Feasible Query
Make = “Toyota”, Model = “Camry”, Price ≤ $7000

Query may be too general
Query may be too specific

What about the price of a Honda Accord?
Is there a Camry for $7100?
**QUIC System** [VLDB 07, ICDE 07, CIDR 07]

**Imprecise Queries**

Query results are no longer exactly satisfying user queries. Rank them based on Expected Relevance Ranking.

\[ \mathcal{ER}(\hat{t}|Q, U, D) = \sum_{t \in C(\hat{t})} \mathcal{R}(t|Q, U) \mathcal{P}(t|\hat{t}, D) \]

**Incomplete Data**

Automated & Non-intrusive assessment of Relevance and Density functions.

Query rewriting to retrieve similar/incomplete tuples in the order of their relevance.
### Challenge: How Should User Believe it?

Query Results for query

\textit{Make like honda and Model like civic and Year like 2001}

<table>
<thead>
<tr>
<th>Make</th>
<th>Model</th>
<th>Year</th>
<th>Price</th>
<th>Mileage</th>
<th>Location</th>
<th>Color</th>
<th>Relevant</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>honda</td>
<td>civic</td>
<td>2001</td>
<td>16662</td>
<td>58977</td>
<td>Tempe</td>
<td>blue</td>
<td></td>
<td></td>
</tr>
<tr>
<td>honda</td>
<td>civic</td>
<td>2001</td>
<td>18610</td>
<td>16667</td>
<td>Mesa</td>
<td>red</td>
<td></td>
<td></td>
</tr>
<tr>
<td>honda</td>
<td>civic</td>
<td>2001</td>
<td>15994</td>
<td>48123</td>
<td>Chandler</td>
<td>silver</td>
<td></td>
<td></td>
</tr>
<tr>
<td>?</td>
<td>civic</td>
<td>2001</td>
<td>13490</td>
<td>58977</td>
<td>Phoenix</td>
<td>silver</td>
<td></td>
<td>This car is 100% likely to have make=honda given that its model=civic</td>
</tr>
<tr>
<td>honda</td>
<td>civic</td>
<td>2003</td>
<td>17490</td>
<td>16667</td>
<td>Phoenix</td>
<td>gray</td>
<td></td>
<td>Cars having year=2003 are 80% similar to cars having year=2001</td>
</tr>
<tr>
<td>honda</td>
<td>accord</td>
<td>2001</td>
<td>15994</td>
<td>48123</td>
<td>Gilbert</td>
<td>silver</td>
<td></td>
<td>Cars having model=accord are 78% similar to cars having model=civic given that 78% of users who looked at civic also looked at accord</td>
</tr>
<tr>
<td>honda</td>
<td>?</td>
<td>2001</td>
<td>14995</td>
<td>32533</td>
<td>Mesa</td>
<td>black</td>
<td></td>
<td>This car is 73% likely to have model=civic given that its make=honda, year=2001, and color=black</td>
</tr>
<tr>
<td>honda</td>
<td>?</td>
<td>2001</td>
<td>15990</td>
<td>43137</td>
<td>Tempe</td>
<td>silver</td>
<td></td>
<td>This car is 32% likely to have model=accord given that its make=honda, year=2001, and color=silver and 78% of users who looked at civic also looked at accord</td>
</tr>
</tbody>
</table>
Discussions

- **A recommendation system:** A query processor that provides results beyond the ones that exactly satisfy a user query.
- **Explanation is important to gain user’s trust.**
- **We need to record data provenance about the reasoning of query processing.**
  - What is provenance information?
    - How the answers are derived (not just computed)?
    - Need to know the reasoning and evidence (knowledge base, statistical information, etc)
  - How should we obtain the provenance information?
    - Proactively recording
  - How should we display the provenance information to users?
    - display (a) in a hierarchical form of different granularities; (b) a chain of provenance step by step
  - Can we query the provenance itself to inspect the system? How?