Entity-Resolution: Beyond The Basics

Steven Whang

Stanford University

Work with: Hector Garcia-Molina, David Menestrina, Omar Benjelloun, Qi Su, Jennifer Widom, Tyson Condie, Nicolas Pombourcq, Tait Larson, Johnson Gong, Makoto Tachibana, Suttipong Thavisomboon, Georgia Koutrika, Martin Theobald, Dave Marmaros
Entity Resolution

N: a  A: b  CC#: c  Ph: e

N: a  Exp: d  Ph: e
Applications

• comparison shopping
• mailing lists
• classified ads
• customer files
• counter-terrorism
ER in the InfoLab

- Pay-As-You-Go ER
- Evolving Rules
- ER Metrics
- Trio-ER
- Joint ER
- Iterative Blocking [SIGMOD 09]
- Negative Rules [VLDB J. 09]
- Generic ER [VLDB J. 09]
- Distributed ER [ICDCS 07]
- Confidences [VLDB CleanDB 06]
Evolving Rules

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
<th>Zip</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>John</td>
<td>54321</td>
<td>123</td>
</tr>
<tr>
<td>r2</td>
<td>John</td>
<td>54321</td>
<td>987</td>
</tr>
<tr>
<td>r3</td>
<td>John</td>
<td>11111</td>
<td>987</td>
</tr>
<tr>
<td>r4</td>
<td>Bob</td>
<td>null</td>
<td>121</td>
</tr>
</tbody>
</table>

N & Z  

r1  r2  r3  r4  (6 comps)
# Evolving Rules

<table>
<thead>
<tr>
<th>Record</th>
<th>Name</th>
<th>Zip</th>
<th>Phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>John</td>
<td>54321</td>
<td>123</td>
</tr>
<tr>
<td>r2</td>
<td>John</td>
<td>54321</td>
<td>987</td>
</tr>
<tr>
<td>r3</td>
<td>John</td>
<td>11111</td>
<td>987</td>
</tr>
<tr>
<td>r4</td>
<td>Bob</td>
<td>null</td>
<td>121</td>
</tr>
</tbody>
</table>

N & Z

N & P

(6 comps)
Rule Evolution

N & Z

N & Z & P

r1  r2  r3  r4
Rule Evolution

N & Z

N & Z & P

(1 comp)
Rule Evolution

N & Z

(1 comp)

N & Z & P

N & P

\( r_1 \ r_2 \ r_3 \ r_4 \)

\( r_1 \ r_2 \ r_3 \ r_4 \)
Materialization

N & Z

N

Z

N & P
Materialization

N & Z
N
Z
N & P

(3 comps)
**ER models**

- Match-based clustering
- Distance-based clustering
Rule Evolution speedup

![Graph showing runtime improvement vs strictness of common comparison rule]

- SN
- HCB
- HCBR
- HCDS
Total Runtime

- SN (Naive)
- SN (Evolve)
- HCB (Naive)
- HCB (Evolve)
- HCBR (Naive)
- HCBR (Evolve)
- HCDS (Naive)
- HCDS (Evolve)
ER in the InfoLab

- Pay-As-You-Go ER
- Evolving Rules
- ER Metrics
- Trio-ER
- Joint ER
- Iterative Blocking [SIGMOD 09]
- Negative Rules [VLDB J. 09]
- Generic ER [VLDB J. 09]
- Distributed ER [ICDCS 07]
- Confidences [VLDB CleanDB 06]