Evaluating Entity Resolution

Results

David Menestrina, Steven Whang, Hector Garcia-Molina

Stanford University
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
Evaluating ER Results

R1 = a, b, c, d, efgh  
R2 = ab, cd, ef, gh  
G = ab, cd, efgh
Pairwise Recall

R1 = a, b, c, d, efgh
R2 = ab, cd, ef, gh
G = ab, cd, efgh
Pairwise Recall

R1 = a, b, c, d, efgh

Pairs:
   ef, eg, eh,
   fg, fh, gh

R2 = ab, cd, ef, gh

Pairs:
   ab, cd, ef, gh

G = ab, cd, efgh

Pairs:
   ab, cd, ef, eg,
   eh, fh, fh, gh
Pairwise Recall

R1 = a, b, c, d, efgh

R2 = ab, cd, ef, gh

G = ab, cd, efgh

Pairs: ef, eg, eh, fg, fh, gh

Pairs: ab, cd, ef, gh

Pairs: ab, cd, ef, eg, eh, fh, fh, gh

6 pairs, all in G

8 pairs
**Pairwise Recall**

R1 = a, b, c, d, efgh  
Pairs:  
  ef, eg, eh,  
  fg, fh, gh  
6 pairs, all in G  
Recall = 6/8 = 75%

R2 = ab, cd, ef, gh  
Pairs:  
  ab, cd, ef, gh

G = ab, cd, efgh  
Pairs:  
  ab, cd, ef, eg,  
  eh, fh, fh, gh  
8 pairs
Pairwise Recall

R1 = a, b, c, d, efgh
Pairs: ef, eg, eh, fg, fh, gh
6 pairs, all in G
Recall = 6/8 = 75%

R2 = ab, cd, ef, gh
Pairs: ab, cd, ef, gh
4 pairs, all in G
Recall = 4/8 = 50%

G = ab, cd, efgh
Pairs: ab, cd, ef, eg, eh, fh, fh, gh
8 pairs

Recall = 6/8 = 75%
Pairwise Recall

R1 = a, b, c, d, efgh

Pairs:
  ef, eg, eh,
  fg, fh, gh

6 pairs, all in G

Recall = 6/8 = 75%

R2 = ab, cd, ef, gh

Pairs:
  ab, cd, ef, gh

4 pairs, all in G

Recall = 4/8 = 50%

G = ab, cd, efgh

Pairs:
  ab, cd, ef, eg,
  eh, fh, fh, gh

8 pairs
**Pairwise F1**

\[
PairPrecision(R, G) = \frac{|Pairs(R) \cap Pairs(G)|}{|Pairs(R)|}
\]

\[
PairRecall(R, G) = \frac{|Pairs(R) \cap Pairs(G)|}{|Pairs(G)|}
\]

\[
pF_1 = \frac{2 \times Precision \times Recall}{Precision + Recall}
\]
Merge Distance

R1 = a, b, c, d, e, f, g, h

R2 = ab, cd, ef, gh

G = ab, cd, e, f, g, h
Merge Distance

\[
\begin{align*}
R_1 &= a, b, c, d, efg h \\
R_2 &= ab, cd, ef, gh \\
G &= ab, cd, efg h
\end{align*}
\]

\[
\begin{align*}
a, b &\rightarrow ab \\
c, d &\rightarrow cd
\end{align*}
\]
Merge Distance

R1 = a, b, c, d, efgh
R2 = ab, cd, ef, gh
G = ab, cd, efgh

a, b → ab

ef, gh → efgh

c, d → cd

G = ab, cd, efgh
G = ab, cd, efgh
Merge Distance

R1 = a, b, c, d, efg

R2 = ab, cd, ef, gh

G = ab, cd, efg

Distance = 1

a, b → ab
c, d → cd

ef, gh → efg

G = ab, cd, efg

Distance = 2

G = ab, cd, efg

Distance = 1
Merge Distance

Minimum number of splits and merges to get from R to G (splits first)
[Al-Kamha, et al. 2004]
Variation of Information

\[ VI(R, G) = H(R) + H(G) - 2I(R, G) \]

\[ H(R) = - \sum_{r \in R} \frac{|r|}{N} \log \frac{|r|}{N} \]

\[ I(R, G) = \sum_{r \in R} \sum_{g \in G} \frac{|r \cap g|}{N} \log \frac{|r \cap g| \times N}{|r| \times |g|} \]
## Conflicts

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pairwise Recall</strong></td>
<td>75%</td>
<td>50%</td>
</tr>
<tr>
<td><strong>Merge Distance</strong></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Variation of Information</strong></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Road Map

Existing Measures

Conflict Evaluation

New Measure

Efficient Measure Algorithm

Runtime Evaluation
Road Map

Existing Measures

Conflict Evaluation

New Measure

Efficient Measure Algorithm

Runtime Evaluation
Generalized Merge Distance

• Cost of split, merge defined by functions:
  \[f_s(x, y), f_m(x, y)\]
  e.g.,
  \[f_m(x, y) = 1\]
  \[f_m(x, y) = xy\]

• Distance = cost of minimum-cost path
Generalized Merge Distance

R1 = a, b, c, d, efg
R2 = ab, cd, ef, gh
G = ab, cd, efg

\[ \begin{align*}
\text{Distance} & = f_m(1, 1) + f_m(1, 1) \\
\text{Distance} & = f_m(2, 2)
\end{align*} \]
\[
f(x, y) = 1
\]

\[\begin{align*}
R1 &= a, b, c, d, efgh \\
R2 &= ab, cd, ef, gh \\
G &= ab, cd, efgh
\end{align*}\]

\[\begin{align*}
a, b &\rightarrow ab \\
c, d &\rightarrow cd \\
ef, gh &\rightarrow efgh
\end{align*}\]

\[\begin{align*}
G &= ab, cd, efgh \\
G &= ab, cd, efgh
\end{align*}\]

\[\begin{align*}
\text{Distance} &= f_m(1, 1) + f_m(1, 1) \\
&= 1 + 1 = 2
\end{align*}\]

\[\begin{align*}
\text{Distance} &= f_m(2, 2) \\
&= 1
\end{align*}\]
\[ f(x, y) = xy \]

\[ R_1 = a, b, c, d, \text{efgh} \]
\[ R_2 = ab, cd, ef, gh \]
\[ G = ab, cd, \text{efgh} \]

\[ a, b \rightarrow ab \]
\[ c, d \rightarrow cd \]
\[ \text{ef, gh} \rightarrow \text{efgh} \]

\[ G = ab, cd, \text{efgh} \]

\[ \text{Distance} = f_m(1, 1) + f_m(1, 1) \]
\[ \text{Distance} = f_m(2, 2) \]
\[ f(x, y) = xy \]

<table>
<thead>
<tr>
<th>R1 = a, b, c, d, efgh</th>
<th>R2 = ab, cd, ef, gh</th>
<th>G = ab, cd, efgh</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, b → ab</td>
<td>ef, gh → efgh</td>
<td></td>
</tr>
<tr>
<td>c, d → cd</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Distance
\[
= f_m(1, 1) + f_m(1, 1) \\
= 1 \times 1 + 1 \times 1 = 2
\]

Distance
\[
= f_m(2, 2) \\
= 2 \times 2 = 4
\]
Relationships Between Measures

- Merge Distance: $f_m(x, y) = 1$, $f_s(x, y) = 1$
- Pairwise Recall: $f_m(x, y) = xy$, $f_s(x, y) = 0$
- Pairwise Precision: $f_m(x, y) = 0$, $f_s(x, y) = xy$
- Variation of Information:
  $$f_m(x, y) = f_s(x, y) = h(x + y) - h(x) - h(y)$$

$$h(x) = \frac{x}{N} \log \frac{x}{N}$$
Road Map

- Existing Measures
- Conflict Evaluation
- New Measure
- Efficient Measure Algorithm
- Runtime Evaluation
Slice Algorithm

- Linear time algorithm
- Extra property required:
  \[ f(x, y) + f(x+y, z) = f(x, z) + f(x+z, y) \]
- Cost functions for pairwise, merge distance, and variation of information (and many others) satisfy property
Slice Algorithm
Slice Algorithm

R

G

Canonical
Road Map

- Existing Measures
  - Conflict Evaluation
  - New Measure
- Efficient Measure Algorithm
  - Runtime Evaluation
Slice Runtime

![Graph showing runtime vs. number of entities for VI (Slice) and VI (straightforward)]
Slice Runtime

![Graph showing runtime vs number of entities for different methods: VI (Slice), VI (straightforward), pF1 (Slice), pF1 (straightforward).]
Conclusion

• Existing measures conflict
• Generalized merge distance provides
  • Configurability to suit different applications
  • Framework for exploring relationships between measures
  • Efficient algorithm (Slice) for computing many distance measures
Thanks!
Relationships Between Measures

- **Merge Distance:** $\text{GMD}(R, G)$
  where $f_m(x, y) = 1$, $f_s(x, y) = 1$

- **Pairwise Recall:** $1 - \frac{\text{GMD}(R, G)}{\text{GMD}(\perp, G)}$
  where $f_m(x, y) = xy$, $f_s(x, y) = 0$

- **Pairwise Precision:** $1 - \frac{\text{GMD}(R, G)}{\text{GMD}(R, \perp)}$
  where $f_m(x, y) = 0$, $f_s(x, y) = xy$

- **Variation of Information:** $\text{GMD}(R, G)$
  where $f_m(x, y) = f_s(x, y) = h(x + y) - h(x) - h(y)$

\[
h(x) = x \log \frac{x}{N}
\]