Human-Powered Blocking in Entity Resolution: A Feasibility Study

Weiling Li, Jongwuk Lee, Dongwon Lee
The Pennsylvania State University
Aug, 2014
A Motivating Example

- **Matching**: the same sport type in an image data set.
- Entity Resolution (ER)
- Challenging!
Machine Based ER Techniques

- **Similarity Based**
  - Might not get an accurate result (e.g., an image data set). 😞

- **Learning Based**
  - Need a good training set to train the classifier. 😞
Crowdsourced ER

- Human workers are assigned tasks referred to the Human Intelligence Tasks (HITs).
- Example:

  Naïve Approach:
  - The # of HITs for ER on $n$ records would be $O(n^2)$. 😞
Crowdsourced ER

- **Blocking:**
  - Group records that are more likely to match into the same “block”. Run pair-wise comparisons only within blocks.

- The # of HITs would be reduced to $O(n + kb^2)$, where $k$ is the number of blocks and $b$ is the average number of records in a block.
Two variations of human-powered blocking:

- An extension of the crowd-median method for blocking [1]
- A hierarchical blocking method
Human-powered Operations

- $\text{hp\_match}(r, r')$

- $\text{hp\_most\_similar}(r_t, C)$
Human-powered Components

- **FindCentroids**
  - Given a data set $D$, the workers choose $K$ centroids.
  - Use $hp\_match(r, r')$

- **Assign**
  - Given a data set $D$, and the set of centroids $C$, the workers assign each record $r$ to one block whose centroid $C_i$ is most similar to $r$.
  - Use $hp\_most\_similar(r_t, C)$

- **PairwiseMatch**
  - Given a data set $D$, the workers make decisions on whether pairs are matched.
  - Use $hp\_match(r, r')$
Human-powered Median-based Blocking

- Human-powered *UpdateCentroids* [1]: Identify a centroid of a data set
  - finding the “outlier” in a set of three records (i.e., triplet).

- block centroid: the least selected one.
- Sampling of the triplets: L=1, H=5
- Crowdsourced k-means clustering: *Assign*, and *UpdateCentroids*.
Human-powered Hierarchical Blocking

- build a K-ary tree of blocks in a top-down fashion
- Split: `FindCentroids` and `Assign`.
- Example:
Stopping criterion:

- $|\text{block}| \leq \text{the pre-defined block size threshold } S$
- The number of HITs due to further blocking ($HIT_b$) exceeds that from direct matching ($HIT_d$).
  - Stringent condition: $\min HIT_b \geq HIT_d$

To improve the overall accuracy, the algorithm runs multiple iterations.
Experimental Setting

- We employ two different HIT designs:
  - binary HIT: processes two records at a time.
  - n-ary HIT: all centroids can be displayed to the workers at once.

Table: # of HITs for each component:

<table>
<thead>
<tr>
<th>Human-Powered Component</th>
<th>Binary HIT</th>
<th>N-ary HIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>FindCentroids</td>
<td>$\geq 1 + 2 + \ldots + (K-1) = \frac{K(K-1)}{2}$</td>
<td>$\geq K-1$</td>
</tr>
<tr>
<td>Assign</td>
<td>$(</td>
<td>D</td>
</tr>
</tbody>
</table>
Evaluation on Synthesis Data

- **Synthesis data**
  - 1000 points.
  - Ground truth: Euclidean distance $d(p, q) \leq a$ distance threshold $T (=1.41)$, points $p$ and $q$ are matched.
  - Parameter setting: $S=100, K=5$. 
Results on Synthesis Data: F1

Baseline (pair-wise matching): F1 = 1.0
Results on Synthesis Data: cost
Evaluation on Real-life Data

- **Image data**
  - 100 images from ImageNet[2].
  - Ground truth: If two images share the same parent node in the hierarchy, they are matched.
- Parameter setting: \( S=15, K=4 \).
- Paid $0.01/question to crowd works on Amazon Mechanical Turk (AMT)
- **Majority voting**
  - Optimization of HIT assignments
This data set has 585 pairs of matching records in eight leaf nodes.
Results on Image Data: F1
Results on Image Data: cost
Conclusions

- Feasibility study for human-powered blocking
  - Relatively High accuracy
  - Much Lower cost compared to the naïve approach
References


Dataset URLs:

http://pike.psu.edu/download/crowdsens14/1k

http://pike.psu.edu/download/crowdsens14/100imgs
Questions?