Robust Entity Resolution Through High Precision Proxy Labels
Chirag Nagpal, Kyle Miller, Benedikt Boecking and Artur Dubrawski
Auton Lab, School of Computer Science, Carnegie Mellon University

Objectives
Development of an entity resolution pipeline to identify sources of advertisements soliciting illegal activities on the internet.

• Utilising strong features like Phone Numbers as a Proxy Ground Truth.
• Training a Binary Classifier as a match function to detect related observations.
• Identifying a match function score to prevent breakdown of ER.
• Developing a heuristic based ‘Blocking Scheme’ ensuring high Precision rate among connected components.
• Applying pipeline to escort ad data from ‘Backpage.com’, to counter human trafficking.

Match Function
To train our ‘match function’ to predict if two observations in our dataset are related, we train a Random Forest binary classifier on Features like:

• Time difference between posting of ads
• Length of longest common substring.
• Special Characters.
• Frequency, Standard Deviation of common Bi-grams, Uni-grams, Images.
• Features extracted from ‘TJBatchExtractor’, a RegEx based Information Extractor, like Hair Color, Price, Location etc.

The Match Function is generated using:

1. Resolving entities with Phone Numbers
2. For every phone number, sampling pairs of connected components.
3. Featurising pairs of components to generate positive and negative samples.
4. Training a Random Forest over the samples using Sci-Kit.[1]

References

Performance

In order to prevent, naive \( n^2 \) comparisons across the entire corpus, we introduce a blocking scheme. The blocking scheme makes use of strong features, to reduce the size of comparisons to be carried out. Thus,

\[
N \leq \sum_{m=1}^{M} n_m^2 \leq N^2
\]

where,

- \( n_m \) is size of block \( m \) and \( m \) is number of blocks.

We use the following features to generate our blocks,

• Images
• Bi-Grams.
• Uni-Grams.

Blocking Scheme

In order to prevent, naive \( n^2 \) comparisons across the entire corpus, we introduce a blocking scheme. The blocking scheme makes use of strong features, to reduce the size of comparisons to be carried out. Thus,

\[
N \leq \sum_{m=1}^{M} n_m^2 \leq N^2
\]

where,

- \( n_m \) is size of block \( m \) and \( m \) is number of blocks.

We use the following features to generate our blocks,

• Images
• Bi-Grams.
• Uni-Grams.

ER over Blocks

Future Work

• Isolating connected components and using rule learning to learn differences.
• Improving scalability of the entire system.
• More robust image analysis, to make better use of the images associated with ads.

Acknowledgements

The authors would like to thank Prof. S.R. Dhore, HoD AIT, University of Pune. We would also like to thank the team from CMU that made the Robotics Institute Summer Scholar’s Program possible.