ACM SIGMOD Programming Contest 2020

Team: PictureMe (Winner)

Mark Blacher Julien Klaus Matthias Mitterreiter

contact: mark.blacher@uni-jena.de

Friedrich-Schiller-University Jena, Germany

Task: Entity Resolution

Identify and match different manifestations of the same real object in a dataset.

Here: around 30,000 camera specs from different e-commerce websites.

Solution Idea

Look at your data!

Part 1: Create Labels (Python)

Grep titles of seven (very clean) websites

- www.canon-europe.com
- www.mypriceindia.com
- www.priceme.co.nz
- www.shopbot.com.au
- www.shopmania.in
- www.ukdigitalcameras.co.uk
- www.wexphotographic.com

More labels are created from all cameras from www.ebay.com that could not be matched so far. Here, fields `brand` and `model` are considered.

Add domain knowledge about different names in different countries for the same camera (e.g., `canon eos 400d` = `canon rebel xti`) from wikipedia.

Part 2: Predict (C++)

Only titles of websites are considered. Check if all items of a label (e.g., `canon eos 1d mark ii`) are present in the title. If so, it is a match. If more than one label is fully present in a title, it is ambiguous and hence, not a match.

Preprocessing:

Clean the data, i.e., remove typos (e.g., `cannon` → `canon`), unify writing (e.g., `cyber-shot`, `cyber shot` → `cybershot`).

Results

precision:	0.99
recall:	0.99
F-measure:	0.99

running time: 0.3 sec.

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) under grant GI-711/5-1.