ACM SIGMOD Programming Contest 2020

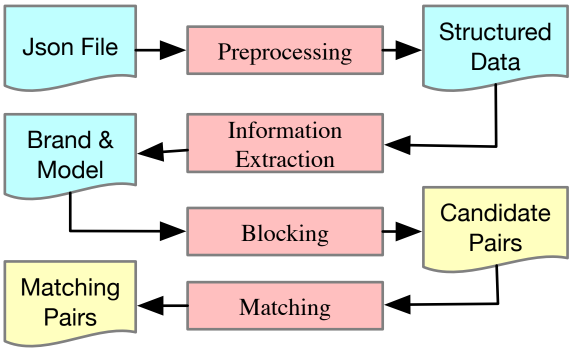
DBTHU Team @ Tsinghua University
Ji Sun, Guoliang Li (advisor)
sun-j16@mails.tsinghua.edu.cn, liguoliang@tsinghua.edu.cn

1. Contest Overview

Task: Identifying which product specifications (in short, specs) from multiple e-commerce websites represent the same real-world product.
Implementation: Python 3.7+ with multiprocessing
Testing Environment: 4 x 3.0 GHz processors, 16 Gb Main Memory, 128 Gb Disk Storage.

2. Framework

Our framework consists of four main components: **Preprocessing module** extracts necessary information from semi-structured json files. **Information Extraction module** extracts precise camera brand and model according to both word semantic and pattern. **Blocking module** groups entities with same brand together. **Matching module** produces matched pairs by rules.



3. Preprocessing

— Filling relational table using attributes in json files including ‘model_id’, ‘brand’, ‘title’, ‘manufacture’, etc.
— Transform letters to lower cases.

4. Information Extracting

Solution Overview: Recognize the precise brand and model for each entity according to the word semantic, context and pattern.
Word semantic: Skip-gram model + human label. E.g., brands have similar embedding.
Word context: Detect types of words according to adjacent landmarks (brand/pixels). E.g. words near brand are likely camera model, version often follows word ‘mark’ (mark III).
Pattern: Detect types of words according to the regex of word. E.g. 600d is apparently a model.

We rank all the extracted models for each entity according to the global tf-idf scores.

5. Blocking & Matching

Blocking: Group all the entities by extracted brand, and conduct Cartesian product in each group to generate candidate pairs.
Matching: Rule-based methods.
Model Matching: If a pair of entities share at least one common model and the model is the first one in either entity, they can be matched.
Version verification: If the model has more versions, version must be matched. E.g. 5dmark2 cannot be matched with 5dmark1.
Synonym: models often co-exist in matched pairs likely be synonyms, and they should be predicted as match. E.g. rebel t3i and eos 600d.

6. Results

Recall: 99%, Precision: 98%, F1: 99%
Latency: < 40 seconds

7. Conclusions

— Entity resolution problems in real world can hardly be killed by one stone.
— Rule-based methods are still effective and efficient when the training data is limited.