ExtER: A High Performance Entity Resolution Framework for e-Commerce Data

SUSTech_DBGroup • ACM SIGMOD Programming Contest 2021 Overall Winner

Weibao Fu fuwb2018@mail.sustech.edu.cn

Peiqi Yin yinpq2018@mail.sustech.edu.cn

Lan Lu 11810935@mail.sustech.edu.cn

Advisor: Bo Tang, Xiao Yan

https://acm.sustech.edu.cn/btang @ DBGroup

Contest Overview

Task: Entity resolution, which identifies pairs of instances (e.g., webpages) that refer to the same object (e.g., notebook).

Input/Output: Three datasets of product descriptions from e-commerce websites/All pairs of matching descriptions in the same dataset.

Performance Metric: F1 score (the harmonic mean of precision and recall).

Solution Framework

Preprocessing: For each instance, extract key features, resolve alias and fill in missing features from its descriptions.

Blocking: Put instances that may refer to the same object into a block according to the features of the instances.

Matching: Enumerate all instance pairs in each block.

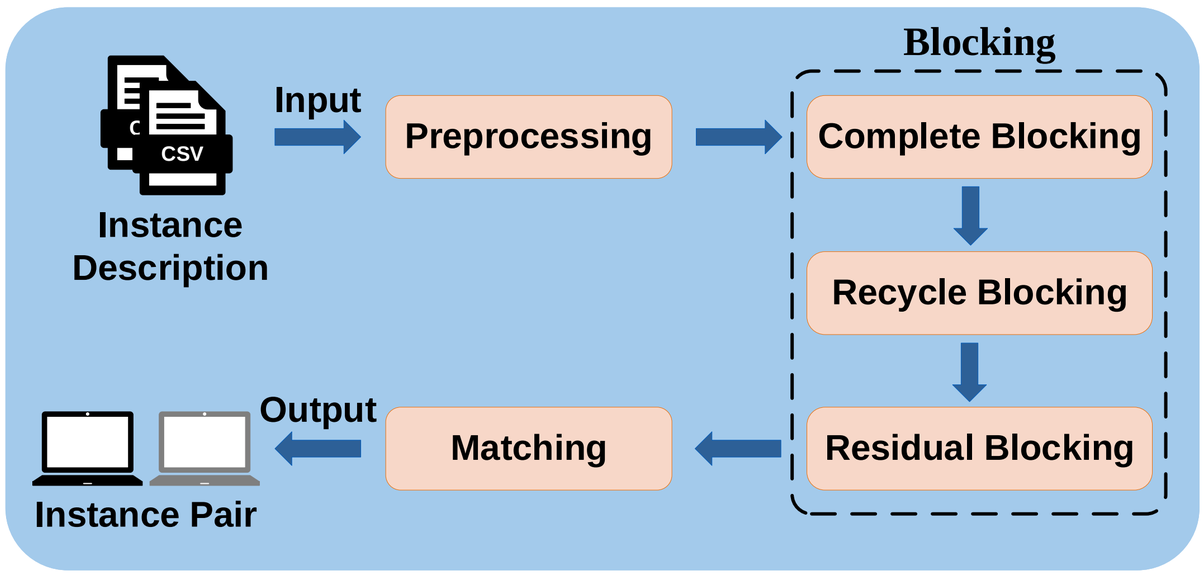

Figure 1: Solution Framework

The core component is ***three-stage blocking***.

Preprocessing

Retrieving Features

- Regular rules to extract key features (e.g., brand and capacity).
  e.g., [0-9]+[\s]?[Gg][Bb] to extract capacity
- Alias tables to resolve aliases.
  e.g., {m620:620m, 620m:620m}

Hashing Index

- Key: Identifier of each instance.
- Value: Address of extracted features structure.

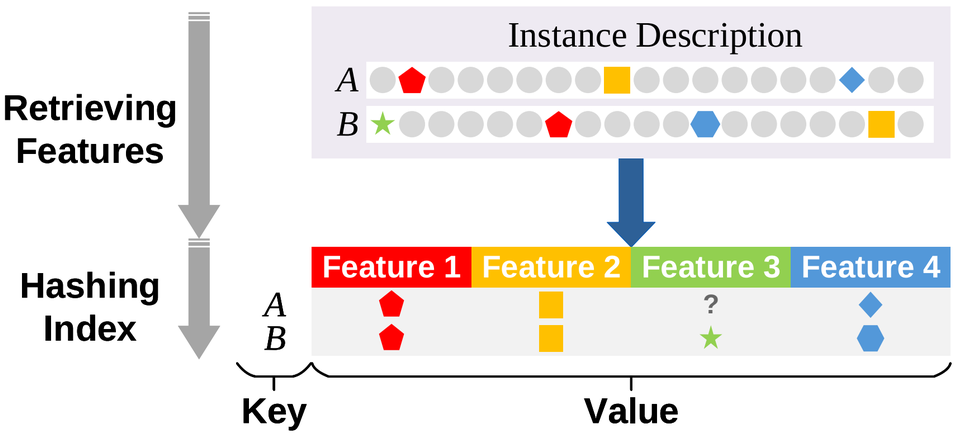

Figure 2: Preprocessing

Blocking

Complete Blocking

<u>Split instances into the solved set and the unsolved set</u>

- Determine primary features for each brand that can uniquely identify an instance. **e.g., (memory, type) for sandisk**
- Put instances that have complete primary features into the ***solved set***, in which each unique combination of the primary features constitutes a block.
- Put other instances into the ***unsolved set***.

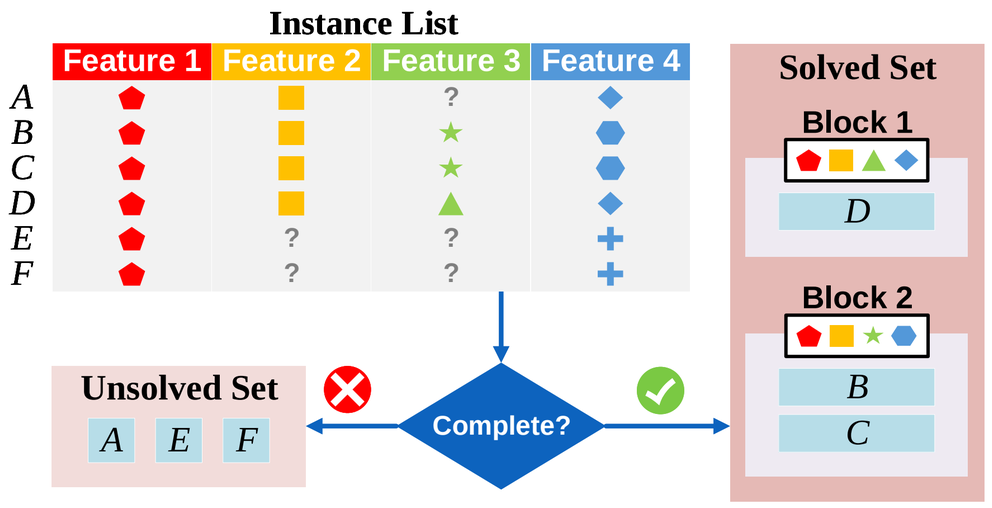

Figure 3: Complete Blocking

Recycle Blocking

<u>Assign instances in the unsolved set</u>

- Design an order to check different combinations of the features.
  e.g., (brand, model) before (brand, memory)
- An instance matches a block if at least one combination of its existent features matches that of the block.
- Each instance is assigned to the first matching block when checking the feature combinations in the predefined order.

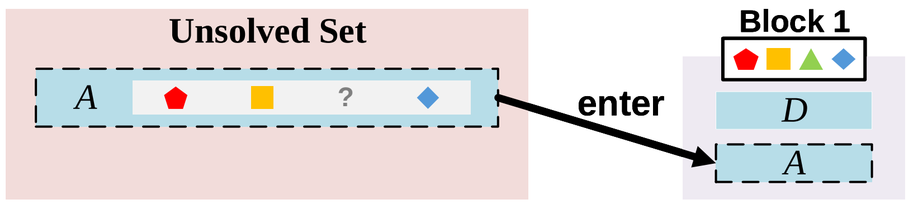

Figure 4: Recycle Blocking

Residual Blocking

<u>Assign instances remained in the unsolved set</u>

- Two instances need to have the same existent and missing features to be assigned to the same block.
  e.g., (toshiba, missing, N401) and (toshiba, missing, N401) are assigned to the same block
- These new blocks are separate from blocks for the solved set.

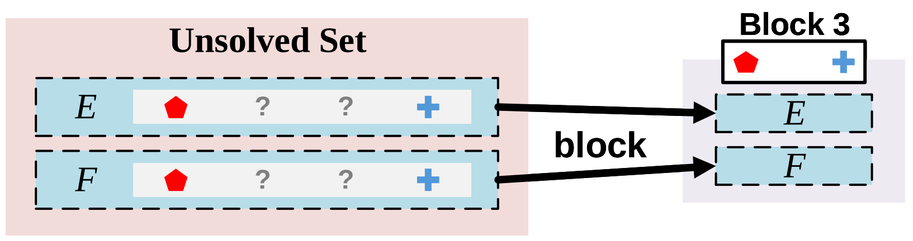

Figure 5: Residual Blocking

Matching

- Enumerate all possible instance pairs for each block.
- A block contains n instances produce $\frac{n(n-1)}{2}$ pairs.

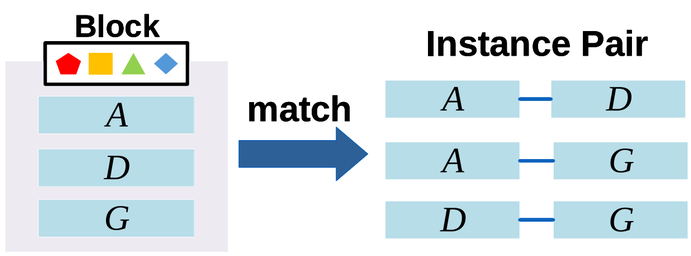

Figure 6: Matching

Result

Dataset	Precision	Recall	F1-Score
X2	0.995	0.971	0.983
X3	0.991	0.980	0.986
X4	0.980	0.880	0.927

Table 1: Result