ACM SIGMOD Programming Contest 2022

Team April

Data Curation Lab - Database Group @ Rutgers University

Chaoji Zuo, Zhizhi Wang Advisor: Dong Deng

chaoji.zuo@rutgers.edu; zw393@rutgers.edu; dong.deng@rutgers.edu

1. Task Overview

Task: Build a **blocking system** for Entity Resolution.

Dataset:

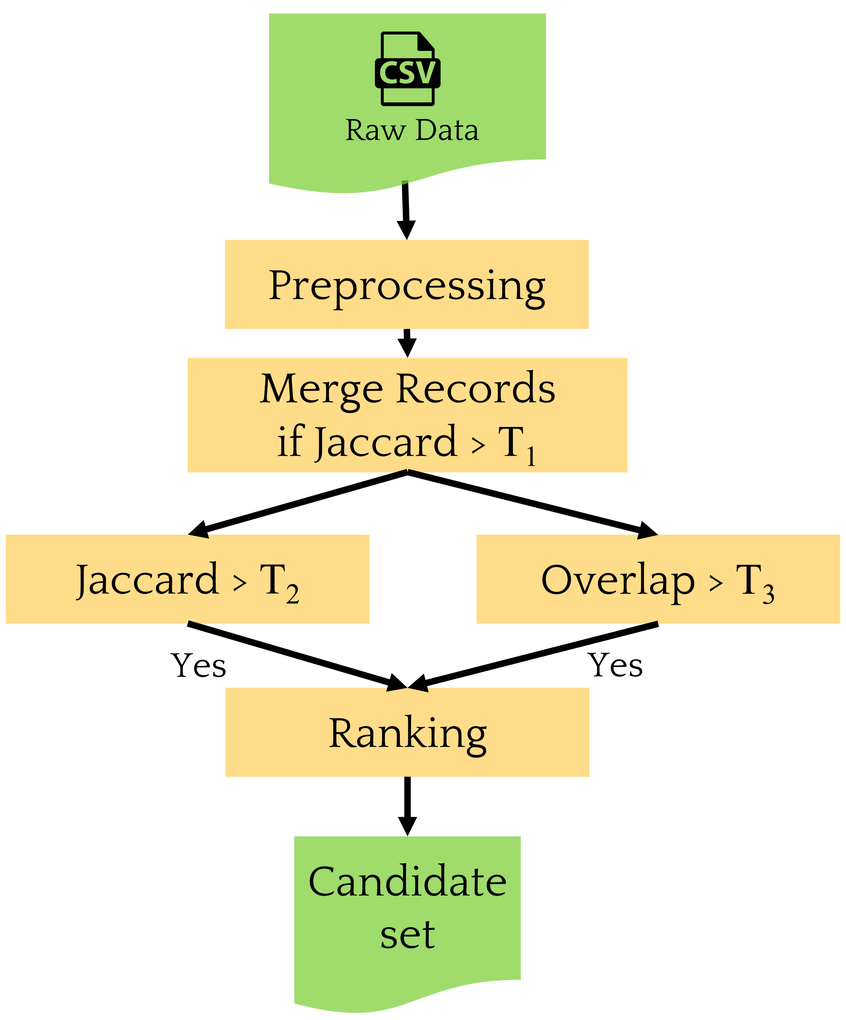
#	Name	# of Rows	Blocking Requirement
Dataset 1	Notebook	1 million	1 million
Dataset 2	Altosight	1 million	2 millions

Measurement:

$$Recall = \frac{\#\ of\ true\ matches\ in\ candidate\ set}{Total\ \#\ of\ true\ matches\ in\ ground\ truth}$$

Evaluation Environment: Azure standard F16s v2 (16 CPU x 2.7 Ghz Processors, 32 Gb Main Memory, 32GB Storage)

2. System Architecture



Our approach leverages **similarity grouping**, **Jaccard join** and **overlap join** to efficiently find a pool of candidate record pairs.

Then, we rank all the candidate record pairs by their weighted similarities and output a fixed number of record pairs as the blocking result.

3. Preprocessing

Using around 10 bash commands to normalize the data:

Data cleaning: Fix typos, drop punctuations, normalization.

A few substitution rules: replace synonyms in different language. (e.g., “mémoire” → “memory”)

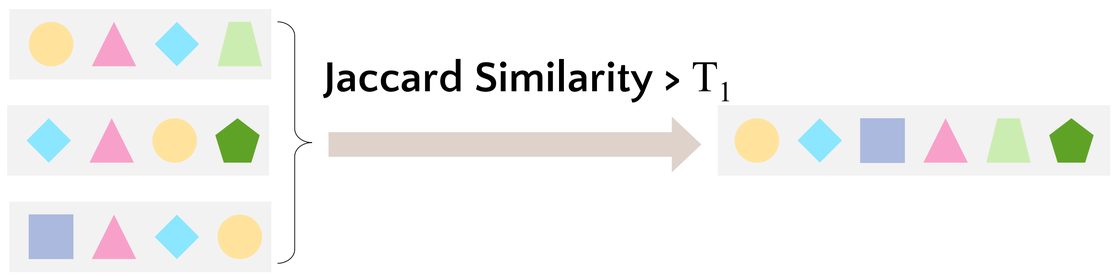
Formatting: Adjust the text to the same format, lowercase letters, replace punctuations. (e.g., “USB 3” → “usb3”)

Tokenization: Tokenize records by whitespace.

4. Deduplication

Observation: many duplicate records in the datasets.

Deduplication:



Goal : make the following similarity join more time efficient

5. Acceleration

1. **Deduplicating** near-duplicate records

   Total number of records reduced.

2. **Parallelization**

   Run the two datasets simultaneously.

   Run the two similarity joins simultaneously.

3. **Efficient similarity join algorithms**

   Adopt the SOTA overlap set join algorithm [1].

   Adopt the SOTA Jaccard similarity join algorithm [2].

6. Result

Average Recall**: 0.520** Runtime: **1679 seconds**

#	T_1	T_2	T_3	Recall
Dataset 1	0.8	0.57	9	0.743
Dataset 2	0.9	0.57	7	0.297

Hindsight:

Could use the training data ground-truth to boost the recall

6. Conclusion & References

- Similarity join is a robust model for blocking that doesn’t rely on heavy data preprocessing or fancy manual rule.
- Parameter setting (adjusting similarity join thresholds) is vital to improve recall, best parameters varies for different dataset.

[1]: Dong Deng, Guoliang Li, He Wen, Jianhua Feng: An Efficient Partition Based Method for Exact Set Similarity Joins. VLDB 2016, Proc. VLDB Endow. 9(4): 360-371 (2015)

[2]: Dong Deng, Yufei Tao, Guoliang Li: Overlap Set Similarity Joins with Theoretical Guarantees. SIGMOD Conference 2018: 905-920