SIGMOD Programming Contest 2021 Entity Resolution

*Saibo Geng, Yifei Li, Lin Yuan, †Panagiotis Sioulas(Advisor)
School of Computer and Communication Sciences, *School of Electrical Engineering, †DIAS Lab
Contact: {saibo.geng, yifei.li, lin.yuan}@epfl.ch

Task Overview

Task: Entity Resolution on datasets of variant domains

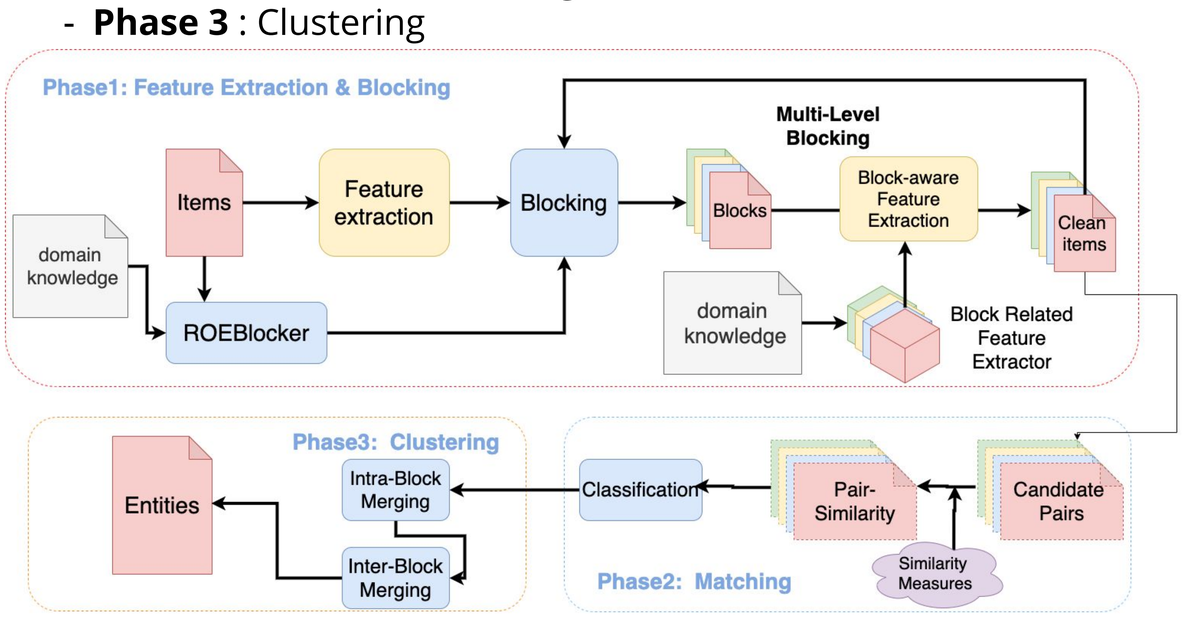
Dataset	# Instance	# entities	# attribs	instance
X2	538	100	14	laptops
X3	605	158	14	laptops
X4*	1356	193	5	SD cards + Flash + SSD

Time Requirements: All pairs of matched instances in 3 datasets in 25 minutes.
Evaluation Metric: Average F-score on the hidden set of 3 datasets
X4* contains multilingualism and products of different nature(SD card vs Flash)

Methodology

Overview

We cast the ER problem to the binary classification problem in the similarity space, such that there are 3 key aspects which determine the system's performance: (a) feature engineering (b) similarity measurement (c) classification
The whole pipeline could be decomposed into 3 phases:
- **Phase 1** : Feature Extraction & Blocking
- **Phase 2** : Instances Matching
- **Phase 3** : Clustering



Phase1 : Feature Extraction & Blocking

A. **General Feature Extraction**:
Input: Tuple of n attributes = >
Output: Tuple of m clean features
B. **Multi-Level Blocking**:
Input: List of tuples = >
Output: List of List of Tuples
Example: Regrouping tuples of laptops by brand into lists of tuples.
C. **Block-Aware Feature Extraction**:
Input: Tuple of m clean features in block X=>
Output: Tuple of m+p_x clean features
D. **Pairing instances**:
input: List of n tuples in block X =>
output: n^2 tuple pairs for block X

Feature Type	Example	Transformation	Similarity Metric
Categorical Feature	CPU Model Brand, CPU type,	Unify name: *Intel Core i5 (3rd Gen) 3320M* => *Core-i5-3320M*	Euclidean Distance
Numerical Feature	Price	Convert various currencies into Euro: 19.99£ => 23.36€	Rounded Absolute Value
String Feature	Product Description	Tokenize,lowercase and truncate to first n tokens	Edit Distance

Phase2 : Instance Matching

E. **Matching**:
Input: pair of two tuples = >
Output: Boolean

Phase3 : Clustering

F. **Clustering**:
Input: m pairs of matching tuples = >
Output: m+n pairs of matching tuples

Strategies:
a. impose transitivity => **TRANS**
b. connect isolated instances with the nearest cluster => **ISO-CON**
c. connect isolated instances which have a have similarity score => **ISO-ISO**

Cluster a **Cluster b** **Cluster c**

Results

Feature Extraction

Feature extraction step demonstrates a strong impact on the performance. We notice the empirical rules are usually easy to craft but much less accurate compared to rules found in products manuals or domain knowledge, but the latter is not always publicly available and can be complex.

Instance Matching

After experimenting with different options for the **matching** and **clustering** phase, we conclude the setting for best performance on each dataset as followed:
- **Notebook** : Model Strict Matching + TRANS
- **Notebook Large**: Random Forest Classifier + ISO-CON
- **Altosight: Product**: Random Forest Classifier + TRANS

Experimental Results

We hereby present the respective performance on each dataset, and the setting we applied. Our final F1-Score is 94.6 on the leaderboard, ranked 4th among all the teams.

Dataset	Matching	Clustering	Precision	Recall	F1-Score
Notebook(X2)	Model Strict Matching	TRANS	93.8	95.3	94.5
Notebook Large(X3)	Random Forest	ISO-CON	99.1	98.9	99.0
Altosight Product(X4)	Random Forest	TRANS	88.0	92.5	90.2

Conclusion & Discussion

Discussion:

Our method performs very well on the two Notebook datasets but less well on **X4.** We believe one important reason is the number of features present on **X4** is much less than **X2** and **X3**. X4 also introduced multilingualism and different currency systems. As the dataset is scraped from real world e-commercial websites, many instances have missing values in some features. This is one notable challenge in this contest.

Conclusion:

This task requires carefully designed rules, either based on empirical observations or from existing conventions. The two strongest challenges are how to tackle **missing values** and how to handle **datasets with few features**.