Université de Paris

New Weighting Schemes for Meta-blocking

Nikolaus Augsten, Roland Kwitt, Matteo Lissandrini, Willi Mann, Themis Palpanas, George Papadakis

LIPADE-TR-Nº 5

October 1, 2021

Technical Report

LIPADE Laboratoire d'Informatique PAris DEscartes

New Weighting Schemes for Meta-blocking

Nikolaus Augsten[1], Roland Kwitt[1],
Matteo Lissandrini[2], Willi Mann[3],
Themis Palpanas[4], George Papadakis[5]

[1]University of Salzburg, Austria {nikolaus.augsten, roland.kwitt}@sbg.ac.at
[2]Aalborg University, Denmark matteo@cs.aau.dk
[3]Celonis GmbH, Germany w.mann@celonis.de
[4]Universite de Paris, France themis@mi.parisdescartes.fr
[5]University of Athens, Greece gpapadis@di.uoa.gr

October 1, 2021

1 Introduction

Entity Resolution constitutes a core data integration task that relies on Blocking in order to tame its quadratic time complexity. Schema-agnostic blocking comes at the cost of many irrelevant candidate pairs (i.e., comparisons), which can be significantly reduced with Meta-blocking. In Meta-blocking, a weighting scheme is first applied on every pair of candidate entities in proportion to the likelihood that they are matching, and a pruning algorithm then discards the pairs with the lowest scores.

In this work, we briefly discuss the existing Meta-blocking weighting schemes, and then propose four new weighting schemes that can be used by Meta-blocking techniques.

2 Existing Meta-blocking Weighting Schemes

The original work on meta-blocking [1] employs a series of *weighting schemes* to assess the co-occurrence patterns of entities in the input block collection. They are all schema-agnostic and produce values that are proportional to the likelihood that the entities in a pairwise comparison are likely to be matching. In the following, with $B_i = \{b \in B \mid e_i \in b\}$ we denote the block set of entity e_i.

1. The *Aggregate Reciprocal Comparisons Scheme* (ARCS) sums the inverse size of the eligible pairs in the common blocks of two entities, i.e., it gives

higher weights to entity pairs that co-occur in smaller blocks:

$$ARCS(e_i, e_j) = \sum_{b_l \in B_i \cap B_j} \frac{1}{|P_l|}$$

2. The *Common Blocks Scheme* (CBS) counts the number of blocks two entities share:

$$CBS(e_i, e_j) = |B_i \cap B_j|$$

3. The *Enhanced Common Blocks Scheme* (ECBS) extends CBS to discount the contribution from entities placed in many blocks:

$$ECBS(e_i, e_j) = |B_i \cap B_j| \cdot \log \frac{|B|}{|B_i|} \cdot \log \frac{|B|}{|B_j|}$$

4. The *Jaccard Scheme* (JS) normalizes the overlap similarity defined by CBS:

$$JS(e_i, e_j) = \frac{|B_i \cap B_j|}{|B_i| + |B_j| - |B_i \cap B_j|}$$

5. The *Enhanced Jaccard Scheme* (EJS) extends JS to discount the contribution of entities that participate in many distinct (i.e., non-redundant) comparisons:

$$EJS(e_i, e_j) = JS(e_i, e_j) \cdot log \frac{|P^1|}{|p^1_i|} \cdot \log \frac{|P^1|}{|p^1_j|}$$

   where P^1 is the set of distinct (i.e., non-redundant) pairs in P, and $p^1_l = \{(e_l, x) \mid (e_l, x) \in P^1\}$ is the set of distinct pairs involving entity e_l.

6. Pearson's χ^2 test extends CBS by assessing whether two adjacent entities e_i and e_j appear independently in the input set of blocks B. To infer their dependency, it estimates whether the distribution of blocks containing e_i in B is the same as the distribution if we exclude the blocks that contain e_j. In more detail, it uses the measures in the contingency Table 1, where $n_{1,1} = |B_i \cap B_j|$ stands for the number of blocks shared by the two entities, $n_{1,2} = |B_i \setminus B_j|$ for the number of blocks containing e_i but not e_j, $n_{2,1} = |B_j \setminus B_i|$ for the number of blocks containing e_j but not e_i and $n_{2,2} = |B \setminus (B_i \cup B_j)|$ for the number of blocks containing neither entity. These are the observed values, whereas the expected value for each cell of the contingency table is $m_{i,j} = \frac{n_{i,+} \cdot n_{+,j}}{n_{+,+}}$. In this context, the edge $e_{i,j}$, which connects e_i and e_j, is weighted according to the following formula: $w_{i,j} = \sum_{i \in \{1,2\}} \sum_{j \in \{1,2\}} \frac{n_{i,j} - m_{i,j}}{m_{i,j}}$.

	$\mathbf{e_j}$	$\neg\mathbf{e_j}$	
$\mathbf{e_i}$	$n_{1,1}$	$n_{1,2}$	$n_{1,+}$
$\neg\mathbf{e_i}$	$n_{2,1}$	$n_{2,2}$	$n_{2,+}$
	$n_{+,1}$	$n_{+,2}$	$n_{+,+}$

Table 1: The contingency table for entities e_i and e_j.

3 New Weighting Schemes

We now propose four new weighting schemes:

1. The *Approximate Enhanced Jaccard Scheme* (AEJS) adapts EJS to a faster functionality, replacing the number of distinct pairs P^1 with the total number of pairs in the input block collection $|P|$ (including the redundant ones) and the number of distinct pairs involving entity e_l, p_l^1, with the total number of pairs in the blocks containing e_l: $p_l = \{(e_l, x) \mid (e_l, x) \in P\}$, including the redundant ones:

$$AEJS(e_i, e_j) = JS(e_i, e_j) \cdot \log \frac{|P|}{|p_l|} \cdot \log \frac{|P|}{|p_l|}.$$

2. The *Weighted Jaccard Scheme* (WJS) essentially normalizes ARCS, i.e., it multiplies every block in the Jaccard coefficient with the inverse size of its eligible pairs:

$$WJS(e_i, e_j) = \frac{\sum_{b_l \in B_i \cap B_j} 1/|P_l|}{\sum_{b_l \in B_i} 1/|P_l| + \sum_{b_l \in B_j} 1/|P_l| - \sum_{b_l \in B_i \cap B_j} 1/|P_l|}.$$

3. The *Reciprocal Sizes Scheme* (RS) is similar to ARCS, but considers the inverse size of common blocks, rather than their eligible pairs:

$$RS(e_i, e_j) = \sum_{b_l \in B_i \cap B_j} \frac{1}{|b_l|}$$

4. The *Normalized Reciprocal Sizes Scheme* (NRS) essentially normalizes RS, i.e., it multiplies every block in the Jaccard coefficient with its inverse size:

$$NRS(e_i, e_j) = \frac{\sum_{b_l \in B_i \cap B_j} 1/|b_l|}{\sum_{b_l \in B_i} 1/|b_l| + \sum_{b_l \in B_j} 1/|b_l| - \sum_{b_l \in B_i \cap B_j} 1/|b_l|}.$$

References

[1] George Papadakis, Georgia Koutrika, Themis Palpanas, Wolfgang Nejdl: Meta-Blocking: Taking Entity Resolutionto the Next Level. IEEE Trans. Knowl. Data Eng. 26(8): 1946-1960 (2014)