Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

REFERENCED FILTERING: A CASE FOR AVOIDING EACH-TO-EACH COMPUTATIONS

Victor Diakov
Tanvi Anandpara

Simfoni Ltd.
450 Townsend St.
San Francisco, CA 94107, USA

ABSTRACT

This study presents a case of avoiding extensive (each-to-each) comparisons in a dataset by using a list of references. Considerations on optimizing the referenced list are formulated, some ensuing results shown.

1 INTRODUCTION

Simulation and optimization of procurements often requires clustering similar supplier names from vendors dataset. In our case, this task is divided into several subtasks with the ultimate goal to combine them under one high-level program, akin to the high-level architecture concept (Dahmann, Fujimoto and Wetherly 1997). The subtasks are: i) cleaning the names dataset of special symbols, stop-words, and duplicates; ii) finding similar names; iii) creating a user-interfaced 'decision engine' to rule which similar names denote the same vendor. The present work focuses on the second subtask from this list.

2 APPROACH

Here, the comparison for two names from the dataset is done by calculating the 'edit distance' between them (https://pupi.org/project/Levenshtein). A 'head-on' approach to finding closely similar names would be to compare all names between them (each-to-each, involves about $N \cdot N / 2$ comparisons), which becomes intractable for large datasets. Our dataset has about 140 thousand entries ($N \sim 140000$) and the 'head-on' approach is impractical.

2.1 Reducing the Number of Calculations

To reduce the number of comparisons, we use reference names. Distance from a reference name to each dataset element (N calculations) helps to exclude most distant pairs from direct distance computation for them. Based on the triangle inequality, the distance between two elements $dist(a,b)$ cannot be below the difference in distances to a third (reference) element r :

$$\text{if} \quad |dist(a,r) - dist(b,r)| \quad \geq t \quad \text{then} \qquad dist(a,b) \geq t,$$

and dataset elements b, whose distance $dist(b,r)$ to the reference r is beyond tolerance t from the reference distance $dist(a,r)$, need not be evaluated for their distance $dist(a,b)$ towards a. Filtering against several references further narrows down the number of candidates for direct evaluation of $dist(a,b)$ (Figure 1).

2.2 Selecting References

A representative subset of about 4 thousand names ($m \sim 4000$) from the initial dataset serves as the starting point for building the reference list. The pair-wise distances $F_{ij} = dist(i,j)$ for the subset are calculated (i and j span the entire 4000 subset). Distances to the reference are intended to serve as filtering criteria, and a

broader range of these distances is expected to be more efficient at filtering. For this reason, the element c of the subset, that shows the highest standard deviation D_c for the distances F_{cj} to the rest of the subset, is selected as the first reference.

The second reference *d*, in addition to showing high standard deviation D_d, should also be dissimilar with the first reference. The absolute value of the correlation C_{cd} between F_{cj} and F_{dj} indicates how 'similar' are *c* and *d*. So the second reference is selected from the subset to maximize $D_d \cdot (1 - |C_{cd}|)$. The rest of the references are selected likewise. A list of 25 references was used in our case to help analyze the initial dataset of ~140 000 names. The references list is partially shown below, with some characters replaced to avoid disclosing possible sensitive data:

```
['QIMTH', 'DSIOFZAOF', 'ZAKAIOF', 'TKTDEMNOADI', 'DNOHBKEIO', ...]
```

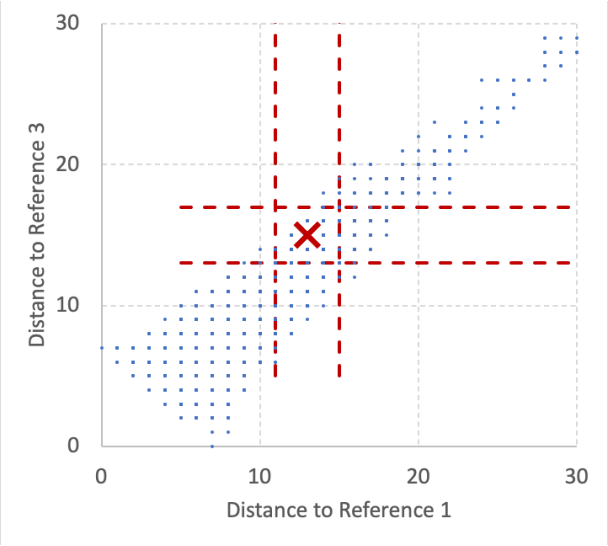

Figure 1: Referenced filtering of dataset elements. For a vendor name (red 'X', at distance 13 from Reference 1 and at 15 from Reference 3), the vendor names represented by blue dots outside the tolerance square (formed by dotted red lines) are guaranteed to be at distance > 2 from the 'red X' vendor, and are excluded from calculating the distance to that element.

ACKNOWLEDGMENTS

The authors gratefully acknowledge Vivek Vibhakar from Simfoni Ltd. for his help with running computer programs on AWS (Amazon Web Services) servers.

REFERENCES

Dahmann, J.S., R.M. Fujimoto, and R.M. Wetherly. 1997. "The Department of Defence High Level Architecture". In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradóttir, K.J. Healy, D.H. Withers, and B.L. Nelson, 142–149. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.