

Scalable N-Way Model Matching Using Multi-Dimensional Search Trees – Summary

Alexander Schultheiß¹ Paul Maximilian Bittner² Thomas Thüm³ Timo Kehrer⁴

Abstract: In this work, we report about recent research on n-way model matching, originally published at the International Conference on Model Driven Engineering Languages and Systems (MODELS) 2021 [Sc21]. Model matching algorithms are used to identify common elements in input models, which is a fundamental precondition for many software engineering tasks, such as merging software variants or views. If there are multiple input models, an n-way matching algorithm that simultaneously processes all models typically produces better results than the sequential application of two-way matching algorithms. However, existing algorithms for n-way matching do not scale well, as the computational effort grows fast in the number of models and their size. We propose a scalable n-way model matching algorithm, which uses multi-dimensional search trees for efficiently finding suitable match candidates through range queries. We implemented our generic algorithm named RaQuN (**R**ange **Q**ueries on **N** input models) in Java, and empirically evaluate the matching quality and runtime performance on several datasets of different origin and model type. Compared to the state-of-the-art, our experimental results show a performance improvement by an order of magnitude, while delivering matching results of better quality.

Keywords: Model-driven engineering; n-way model matching; clone-and-own development; software product lines; multi-view integration; variability mining

1 Summary

Matching algorithms are an essential requirement for detecting common parts of development artifacts in many software engineering activities. In domains where model-driven development has been adopted in practice, such as automotive, avionics, and automation engineering, numerous model variants emerge from cloning existing models. Integrating such autonomous variants into a centrally managed software product line in extractive software product-line engineering requires to detect similarities and differences between them, which in turn requires to match the corresponding model elements of the variants.

Currently, almost all existing matching algorithms can only process *two* development artifacts, whereas it is typically required to identify corresponding elements in *multiple* (i.e., $n > 2$) input models. A few approaches calculate an n-way matching by repeated

¹ Humboldt-University of Berlin, Germany, alexander.schultheiss@informatik.hu-berlin.de

² University of Ulm, Germany, paul.bittner@uni-ulm.de

³ University of Ulm, Germany, thomas.thuem@uni-ulm.de

⁴ Humboldt-University of Berlin, Germany, timo.kehrer@hu-berlin.de

two-way matching of the input artifacts. In each step, the resulting two-way correspondences are simply linked together. However, sequential two-way matching of models may yield sub-optimal results because not all input artifacts are considered at the same time and order in which input models are processed influences the quality of the matching.

The only matching approach which simultaneously processes n input models is a heuristic algorithm called NwM by Rubin and Chechik. Yet, we faced scalability problems when applying NwM to models of realistic size, comprising thousands of elements. The most likely reason for this is the huge number of model element comparisons, which often leads to performance problems even in the case of few input models if these models are large.

We propose RaQuN (**R**ange **Q**ueries on **N** input models), a generic, heuristic n -way model matching algorithm. The key idea behind RaQuN is to map the elements of all input models to points in a numerical vector space. RaQuN embeds a multi-dimensional search tree into this vector space to efficiently find nearest neighbors of elements. By comparing an element only with its nearest neighbors serving as match candidates, RaQuN can reduce the number of required comparisons considerably.

For our empirical assessment, we use datasets from different domains and development scenarios. Next to academic and synthetic models, we investigate variants generated from model-based product lines, and reverse-engineered models from clone-and-own development. Moreover, we are the first to provide a thorough investigation of n -way model matching on large-scale subjects and a real-world clone-and-own subject. Our evaluation shows that RaQuN reduces the number of required comparisons by more than 90% for most experimental subjects, making it possible to match models of realistic size simultaneously.

2 Data Availability

The original publication is accessible under the DOI 10.1109/MODELS50736.2021.00010. A preprint is also available online <https://github.com/SoftVarE-Group/Papers/blob/master/2021/2021-MODELS-Schultheiss.pdf>. Our artifact is available on Github (<https://github.com/AlexanderSchultheiss/RaQuN>) and Zenodo (DOI: 10.5281/zenodo.5150388).

3 Acknowledgments

We thank Lars Grunske for his ongoing support and initiative towards our original publication.

Bibliography

- [Sc21] Schultheiß, Alexander; Bittner, Paul Maximilian; Grunske, Lars; Thüm, Thomas; Kehrer, Timo: Scalable N-Way Model Matching Using Multi-Dimensional Search Trees. In: Proc. Int'l Conf. on Model Driven Engineering Languages and Systems (MODELS). IEEE, Washington, DC, USA, pp. 1–12, October 2021.