A Short Survey on Process Model Similarity

R.M. Dijkman, B.F. van Dongen, M. Dumas, L. García-Bañuelos, M. Kunze,
H. Leopold, J. Mendling, R. Uba, M. Weidlich, M. Weske, Z. Yan

Abstract Process model similarity has developed into a prolific field of investigation. This paper summarizes the research after the CAISE 2008 paper on this topic. We identify categories of problems and provide an outlook on future directions.

1 Introduction

Analysing the similarity of process models has become a dynamic field of research in business process management. This short paper serves as commentary to the CAISE 2008 paper on "Measuring Similarity between Business Process Models" [1] - one of the first major conference papers on this topic after early papers by Mendling, van Dongen and van der Aalst [2], Ehrig, Koschmider and Oberweis [3], Eshuis and Grefen [4], and Corrales, Grigori and Bouzeghoub [5] were published before. The article by Dijkman et al. [6] summarizes contributions before 2011.

Remco Dijkman, Boudewijn van Dongen, Zhiqiang Yan
TU Eindhoven, 5600 MB Eindhoven, The Netherlands e-mail: R.M.Dijkman@tue.nl; b.f.v.dongen@tue.nl; zhiqiang.yan.1983@gmail.com

Marlon Dumas, Luciano García-Bañuelos, Reina Uba
University of Tartu, 50409 Tartu, Estonia e-mail: marlon.dumas@ut.ee; luciano.garcia@ut.ee; reinak@ut.ee

Matthias Kunze, Mathias Weske
HPI – University of Potsdam, 14482 Potsdam, Germany e-mail: matthias.kunze@hpi.uni-potsdam.de; mathias.weske@hpi.uni-potsdam.de

Henrik Leopold
Humboldt-University Berlin, 10099 Berlin, Germany e-mail: henrik.leopold@wiwi.hu-berlin.de

Jan Mendling
Wirtschaftsuniversität Wien, 1090 Wien, Austria e-mail: jan.mendling@wu.ac.at

Matthias Weidlich
Technion – Israel Institute of Technology, 32000 Haifa, Israel e-mail: weidlich@tx.technion.ac.il

The aim of this paper is to summarize the essential directions that emerged from these initial papers. Section 2 discusses the similarity problem in a very general way. Section 3 reviews alternative notions for calculating the similarity of process models. Section 4 turns to the problem of finding matching activities in a pair of process models. Section 5 highlights how the calculation of similarity between process models can help in search and indexing. Section 6 identifies directions of future research as a conclusion of this commentary.

2 The Challenge of Process Model Similarity

Process model similarity calculation is hindered by multiple inherent sources of heterogeneity. Even if two process models define exactly the same behaviour at the same level of granularity and with the same projection on the real-world process, the process models might still look quite different. We might encounter heterogeneity of behavioural representation, labelling styles, and terminology [7].

The first reason for this observation is that the representation of the same behaviour can be achieved with different structures. Partially, this phenomenon relates to the option to "multiply out" different choices in the process. Indeed, corresponding techniques are defined in [8] for making an unstructured process model structured. Therefore, we cannot assume that a different structure of process models does actually imply a difference in behaviour. Second, it has been observed that the labels can be formulated in different grammatical ways. Activities like "Send Invoice" (verb plus object) and "Sending of Invoice" (gerund plus preposition plus object) clearly point to the same type of activity. However, we cannot assume that a difference in the grammatical structure implies a difference in the activity. Techniques for automatically transforming activity labels to a canonical verb-object style are presented in [9]. Third, we can use syntactically different terms to defer to the same matter. For instance, two activity labels in verb-object style like "Check Invoice" and "Evaluate Bill" might use synonymous terms to refer to the same matter. Here again, we cannot assume that a difference in terms always implies a difference in meaning. Calculating the similarity of process models becomes much easier in a setting where we can assume that these heterogeneities are resolved.

These issues of heterogeneous representation can be present even if two models capture one process at the same level of granularity and using the same projections. Yet, even projections and granularity may differ. In the first case, we have to deal with problems that parts of a first process model might simply be left out in a second process model. The question then becomes to which extent the calculation of similarity should punish such a difference in projection. Technically this question relates to the properties of underlying notions for similarity calculation. For the second case, we have to consider questions of granularity. In terms of similarity, it has to be decided in how far a sequence of activities in one model shall be punished when it is shown as a single abstracted activity in the second model.

3 Underlying Notions for Process Model Similarity Calculation

Most approaches for process similarity are based on either the process model's graph structure or the behaviour captured in the model [10]. Similarity is typically quantified by symmetric and non-negative distance functions that capture the amount of differences a pair of process models exposes. Accordingly, two process models are identical if their distance equals 0.

The process model graph plus the execution semantics of its elements prescribe the allowed behaviour of the process, which is typically analysed by means of reachability graphs or the set of allowed execution traces. The problem of calculating the similarity of two process models based on both these options is that both approaches suffer from exponential complexity due to concurrency and loops in the process model [11]. Hence, abstractions have been proposed that only consider the order in which two activities can be executed in any process instance [12]. Two variants are (1) the transition adjacency relation [13, 14] that consider pairs of activities that can be executed directly after each other (non-transitive) and (2) weak order relations [4, 15] which consider any pair of activities that can be executed after each other eventually (transitive). An extension of the latter are behavioural profiles, which distinguish these relations by mutual exclusion, strict, and interleaving order [16, 17]. Although these relations abstract from certain behavioural aspects, e.g., causality and cardinality, they have been shown to support the human assessment of process model similarity [18]. Higher precision can be achieved based on event structures which yield a matrix of relations that fully characterizes a model in terms of a strong notion of behavioral equivalence [19].

An alternative to a notion of distance based on behavioural relations is graph edit distance [20]. The graph edit distance is the minimum number of basic graph operations that is needed to transform one graph into another. The basic operations are: add node, remove node, add edge and remove edge. In labelled graphs, node substitution can also be used as an operation, in which one node is substituted by another node with a different label. The graph edit distance can be transformed into a similarity metric in different ways, e.g., by dividing the distance by the number of nodes and edges of the largest graph and using one minus the result of that.

4 Process Model Matching

A basic technique required for many approaches to process model similarity is matching, the construction of correspondences between the process model activities. Process model matching is inspired by schema and ontology matching [21, 22] and adopts techniques for syntactic or semantic matching proposed in these fields. Despite the conceptual similarities, the problem of matching process models differs from the one of matching data schemas. For instance, the distinguished labelling styles observed for activities and the execution semantics of a process model may

be leveraged for matching process models. On other other hand, unlike in schema matching, instance data is typically not available for matching.

Recently, several approaches for process model matching have been presented. Most of them combine concepts for textual comparison of activity labels with the aforementioned similarity measures. A generic architecture for process model matching is defined by the ICoP framework [23]. Following this architecture, a matcher may rely on the string edit distance for comparing activity labels and a structural similarity measure for process model graphs, as presented in [24]. Activity labels have also been compared based on semantic annotations derived by part of speech tagging. Leopold et al. [25] derive match hypotheses from these annotations and rely on probabilistic inference for the construction of correspondences.

A major challenge for process model matching are differences in modelling granularity. The construction of complex 1:n or even n:m correspondences between activities is hindered by a combinatorial problem: there are exponentially many activity subsets in either process model that form possible candidate correspondences. Heuristics to select candidate correspondences are based on the graph distance [23] or structural decomposition of the process model graph [23, 26]. Then, sets of activities (potentially including their descriptions) are textually compared using coefficients over terms or bigrams [26] or vector space scoring [23].

5 Process Model Search and Indexing

One of the applications of process similarity lies in finding all process models in a process model collection that are sufficiently similar to a so-called “search process model”. Indexing techniques are required for implementing such a search efficiently. Pairwise computation of the distance of process models allows comparing a given query with models from a process repository, and thus, to find similar models. They also need to be ranked [27]. However, traditional indexes cannot be applied as the notion of pairwise similarity does not yield any ordering of process models.There are two competing approaches to indexing process models: (1) indexing process model elements [28, 29]; and (2) indexing complete process models, provided a similarity metric is used that satisfied the triangle inequality property [30].

The first approach is based on breaking each process model up into parts on which existing indexing techniques can be applied. In particular MTree [28] and B+ tree [29] indexing techniques have been used. The types of parts that have been used include: the labels of the tasks in the process models, paths of subsequent tasks and more complex constructs such as choices between tasks or parallel tasks. Using these indexing techniques, the search for similar models is executed by breaking the search model up into parts (e.g. only selecting the task labels) and then using the index to find the process models that have sufficiently many similar parts (e.g. similar task labels). The work on process model search and indexing is related to general graph search and its application to e.g. face recognition and fingerprint search.

The second approach uses metric spaces [31] to tackle this problem, as they facilitate efficient search in the absence of coordinates and ordering of elements if the distance function, beside the properties mentioned above, satisfies the triangle inequality. This property allows determining minimum and maximum distance of two process models without computing the distance function, if their pairwise distance to a third model is given. Such transitivity improves search efficiency and can be applied to both structural [32] and behavioural [18] process model similarities.

6 Future Research on Process Model Similarity

Though we have seen exciting advancements in research on process model similarity, several challenges still await solutions. While behavioural and label styles can be homogenized with recent techniques, we are missing approaches to harmonize terminology and the level of granularity. These challenges are specifically important for process model matching. Another open issue is the lack of reference samples to perform thorough evaluations of different approaches. Dijkman et al. [6] conducted a survey and let process model experts judge on the similarity of 1000 pairs of process models, which helped to refine techniques. More such datasets are required in order to increase the degree of repeatability and reproducibility in this field of research. This will provide a good basis for further comparative studies such as [33]. Related application scenarios such as clustering process models [34], detection of clones [35] and behavioural patterns [36] are expected to further benefit from research into process model similarity.

References

1. van Dongen, B.F., Dijkman, R.M., Mendling, J.: Measuring similarity between business process models. In Bellahsene, Z., Léonard, M., eds.: CAiSE 2008. LNCS 5074 (2008) 450–464
2. Mendling, J., van Dongen, B.F., van der Aalst, W.M.P.: On the degree of behavioral similarity between business process models. In Nüttgens, M., Rump, F.J., Gadatsch, A., eds.: EPK Workshop 2007. Volume 303 of CEUR Workshop Proceedings, CEUR-WS.org (2007) 39–58
3. Ehrig, M., Koschmider, A., Oberweis, A.: Measuring similarity between semantic business process models. In Roddick, J.F., Hinze, A., eds.: APCCM 2007. CRPIT 67 (2007) 71–80
4. Eshuis, R., Grefen, P.W.P.J.: Structural matching of bpel processes. In: ECOWS 2007, IEEE Computer Society (2007) 171–180
5. Corrales, J.C., Grigori, D. Bouzeghoub, M.: BPEL Processes Matchmaking for Service Discovery. In: OTM 2006, Part I, LNCS 4275 (2006) 237–254
6. Dijkman, R.M., Dumas, M., van Dongen, B.F., Käärik, R., Mendling, J.: Similarity of business process models: Metrics and evaluation. Inf. Syst. **36**(2) (2011) 498–516
7. Mendling, J.: Three challenges for process model reuse. In Daniel, F., Barkaoui, K., Dustdar, S., eds.: BPM Workshops (2). LNBIP 100 (2011) 285–288
8. Polyvyanyy, A., García-Bañuelos, L., Dumas, M.: Structuring acyclic process models. Inf. Syst. **37**(6) (2012) 518–538
9. Leopold, H., Smirnov, S., Mendling, J.: On the refactoring of activity labels in business process models. Inf. Syst. **37**(5) (2012) 443–459

10. Dumas, M., García-Bañuelos, L., Dijkman, R.M.: Similarity search of business process models. IEEE Data Eng. Bull. **32**(3) (2009) 23–28
11. Valmari, A.: The state explosion problem. In Reisig, W., Rozenberg, G., eds.: Petri Nets. LNCS 1491 (1996) 429–528
12. van der Aalst, W.M.P., Weijters, T., Maruster, L.: Workflow mining: Discovering process models from event logs. IEEE Trans. Knowl. Data Eng. **16**(9) (2004) 1128–1142
13. Bae, J., Liu, L., Caverlee, J., Zhang, L.J., Bae, H.: Development of distance measures for process mining, discovery and integration. Int. J. Web Service Res. **4**(4) (2007) 1–17
14. Zha, H., Wang, J., Wen, L., Wang, C., Sun, J.: A workflow net similarity measure based on transition adjacency relations. Computers in Industry **61**(5) (2010) 463–471
15. Weidlich, M., Mendling, J., Weske, M.: A foundational approach for managing process variability. In Mouratidis, H., Rolland, C., eds.: CAiSE 2011. LNCS 6741 (2011) 267–282
16. Weidlich, M., Mendling, J., Weske, M.: Efficient consistency measurement based on behavioral profiles of process models. IEEE Trans. Software Eng. **37**(3) (2011) 410–429
17. Weidlich, M., Polyvyanyy, A., Mendling, J., Weske, M.: Causal behavioural profiles - efficient computation, applications, and evaluation. Fundam. Inform. **113**(3-4) (2011) 399–435
18. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity - a proper metric. In Rinderle-Ma, S., Toumani, F., Wolf, K., eds.: BPM 2011. LNCS 6896 (2011) 166–181
19. Armas-Cervantes, A., García-Bañuelos, L., Dumas, M.: Event structures as a foundation for process model differencing, part 1: acyclic processes, In: WS-FM 2012. (to appear)
20. Dijkman, R.M., Dumas, M., García-Bañuelos, L.: Graph matching algorithms for business process model similarity search. In Dayal, U., Eder, J., Koehler, J., Reijers, H.A., eds.: BPM 2009. LNCS 5701 (2009) 48–63
21. Bellahsene, Z., Bonifati, A., Rahm, E., eds.: Schema Matching and Mapping. Springer (2011)
22. Euzenat, J., Shvaiko, P.: Ontology matching. Springer (2007)
23. Weidlich, M., Dijkman, R.M., Mendling, J.: The ICoP framework: Identification of correspondences between process models. In CAiSE 2010. LNCS 6051 (2010) 483–498
24. Dijkman, R.M., Dumas, M., García-Bañuelos, L., Käärik, R.: Aligning business process models. In: EDOC 2009, IEEE Computer Society (2009) 45–53
25. Leopold, H., Niepert, M., Weidlich, M., Mendling, J., Dijkman, R.M., Stuckenschmidt, H.: Probabilistic optimization of semantic process model matching. In Barros, A.P., Gal, A., Kindler, E., eds.: BPM 2012. LNCS 7481 (2012) 319–334
26. Branco, M.C., Troya, J., Czarnecki, K., Küster, J.M., Völzer, H.: Matching business process workflows across abstraction levels. In France, R.B., Kazmeier, J., Breu, R., Atkinson, C., eds.: MoDELS 2012. LNCS 7590 (2012) 626–641
27. Grigori, D., Corrales, J.C., Bouzeghoub, M., Gater, A.: Ranking bpel processes for service discovery. IEEE T. Services Computing **3**(3) (2010) 178–192
28. Yan, Z., Dijkman, R., Grefen, P.: Fast business process similarity search. Distributed and Parallel Databases **30** (2012) 105–144
29. Jin, T., Wang, J., Wu, N., La Rosa, M., ter Hofstede, A.: Efficient and accurate retrieval of business process models through indexing. In: OTM, Part I. LNCS 6426 (2010) 402–409
30. Kunze, M., Weidlich, M., Weske, M.: Behavioral similarity – a proper metric. In: BPM 2011. LNCS 7481 (2011) 166–181
31. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search - The Metric Space Approach. Volume 32 of Advances in Database Systems. Kluwer (2006)
32. Kunze, M., Weske, M.: Metric trees for efficient similarity search in large process model repositories. In zur Muehlen, M., Su, J., eds.: BPM Workshops. LNBIP 66 (2010) 535–546
33. Becker, M., Laue, R.: A comparative survey of business process similarity measures. Computers in Industry **63**(2) (2012) 148–167
34. Niemann, M., Siebenhaar, M., Schulte, S., Steinmetz, R.: Comparison and retrieval of process models using related cluster pairs. Computers in Industry **63**(2) (2012) 168–180
35. Ekanayake, C.C., Dumas, M., García-Bañuelos, L., Rosa, M.L., ter Hofstede, A.H.M.: Approximate clone detection in repositories of business process models. In Barros, A.P., Gal, A., Kindler, E., eds.: BPM 2012. LNCS 7481 (2012) 302–318
36. Smirnov, S., Weidlich, M., Mendling, J., Weske, M.: Action patterns in business process model repositories. Computers in Industry **63**(2) (2012) 98–111