On the notion of diversity in the context of information aggregation in BPM

Clemens Schreiber [1]

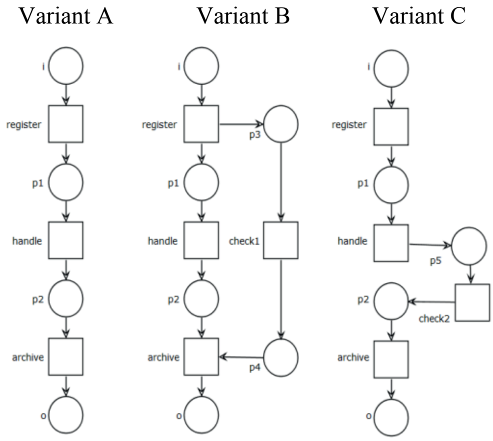
Abstract: Assuming a set of diverse process variants, information aggregation is used to provide an aggregated overview of the overall process. The main challenge is to provide a condensed but accurate overview, while avoiding redundant information. Although there exist different approaches for information aggregation, none of them actually consider the diversity of the multiple process variant attributes, such as resources, cost or duration. But this hinders from gaining an overview of the process diversity, i.e. the variance of multiple process attributes. Hence, valuable information regarding resource and time variability can get lost. Diversity measures can be used to evaluate the significance of a single process variant, based on multiple attributes. This short paper introduces existing diversity measures from different research disciplines and elaborates on their application in BPM for the purpose of information aggregation.

Keywords: Business process variants; Process diversity; Delta analysis

1 Introduction

The problem of providing an aggregated overview of a set of process variants was first described by van der Aalst and Basten [Aa02] as “management-Information problem”. It refers to the challenge of providing a condensed but accurate overview of multiple process variants, while avoiding redundant or trivial information. In [Aa02] a variant or process version is defined as a workflow-net derived from another workflow-net based on an ad-hoc or an evolutionary change of structural nature. Since it is not desirable to show multiple variants of a process individually, one needs to create an aggregated overview of these variants, e.g. in the form of a unique process diagram. The sufficiency and accuracy of the information provided by the aggregated overview depends, however, on the diversity of the process, i.e. the variance of the process variant attributes. The attributes can be manifold and either refer to the control flow (e.g., number of activities, connectivity of the activities) or performance (e.g., case frequency, task durations, output quality). The more diverse two process variants are, the more information should an aggregated overview contain about these variants. An incomplete process overview could for example omit valuable information regarding the overall flexibility of a process or execution paths, which are relevant for the overall performance. Process diversity should therefore always be considered with respect to performance. Another important aspect of the process diversity is that interdependencies between different process variants might become more apparent, due to the correlation between certain process attributes.

[1] Karlsruher Institut für Technologie, clemens.schreiber@kit.edu



Event Log Variant A

ID	Event	Time	Resource	Cost
1	r	8:00-8:10	Employee1	0
1	h	8:15-8:45	Employee2	20
1	a	8:50-9:00	Employee3	0

Event Log Variant B

ID	Event	Time	Resource	Cost
1	r	8:00-8:10	Employee1	0
1	c	8:15-8:30	Employee4	5
1	h	8:15-8:45	Employee2	10
1	a	8:45-9:00	Employee3	0

Event Log Variant C

ID	Event	Time	Resource	Cost
1	r	8:00-8:10	Employee1	0
1	h	8:15-8:45	Employee2	20
1	c	8:45-8:50	Employee2	0
1	a	8:50-9:00	Employee3	0

Fig. 1: Three process variants taken from [Aa02] with three respective event logs

In order to be able to consider a variety of process attributes for information aggregation, one should consider the predefined process model, as well as the execution data. Figure 1 shows three process variants taken from [Aa02]. To emphasize the importance of process attributes beyond the control-flow, the three given event logs show possible process execution data for each respective process variant. These event logs contain information regarding resources, event duration, cost and case frequency. One can see that process variant 2 contributes significantly to the overall process diversity, since 1 different resource is used and the costs are also differently distributed compared to the other variants. The two variants A and C on the other hand are much more similar. This should imply, that compared to variant A, variant B is more important than variant C for the representation of the process diversity and hence, for a more complete process overview.

Figure 2 shows two possible representations of the three process variants in one diagram, also taken from [Aa02]. The first one is based on the concept of the Greatest Common Divisor (GCD) and the second one is based on the Least Common Multiple (LCM). The GCD represents the behaviour that all variants agree upon. The Least Common Multiple is the most compact representation, which is still a subclass of all variants. For a more detailed description on how the two representations are derived see [Aa02]. The LCM does provide significantly less information regarding the overall process diversity as the GCD. Especially when one considers the process attributes costs and resources from variant B. The GCD on the other hand does not depict the execution path <r,h,a>, which could be important if the flexibility of the control flow is required as management information. Hence, this small example shows that process attributes can differ in the importance for the overall process diversity, depending on the purpose of the diversity measure. Several different diversity measures exist, which are applied in different research disciplines such as biology and economics. They each have different properties and allow for different interpretations of diversity. In the following section, several diversity measures will be introduced and analysed in the context of information aggregation in BPM. The purpose of the diversity measure is to identify the process variants that are most significant for the representation of the overall process.

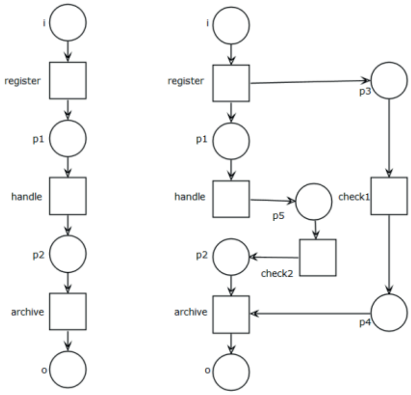


Fig. 2: The LCM (left) and the GCD (right) of the three variants in Fig. 1, taken from [Aa02]

2 Diversity Measures

Diversity can generally be defined as the totality of the dissimilarity among objects [Ne09]. Hence, the more dissimilar objects are among each other, the more diverse is their totality. Biodiversity for example aims to capture the overall dissimilarity of species, which live in a defined space. In economics diversity can refer to the overall dissimilarity among a set of products, such as the model portfolio of a car brand. However, in both cases different aspects of diversity are considered. While for biodiversity the relative occurrence of an object is important, in economics it is the characteristic features of the objects. This can be linked to a normative judgement on the purpose of diversity in its particular domain [Ba06]. In economics, product diversity is usually linked to the ability to choose between different options, in biology, biodiversity is used as an indicator for the preservation of an ecosystem. In the following 3 main approaches for the measurement of diversity among objects will be considered: entropy measures, distance measures and attribute measures. One should notice that further measures exist to capture the diversity across populations and within an object, i.e. the variance of an attribute.

Entropy measures are commonly used for the assessment of biodiversity. They can be considered as special cases of the generalized function [Ba06]:

$$f(n,p) = (\sum \quad p \quad — \tag{1}$$

where is the total number of different species and the relative abundance of a species }. The absolute abundance of a species is therefore represented by a_i. The parameter α determines how much weight the diversity measure places on the abundance of the species. For α = 0, the equation yields n, the total number of different species. In this case the relative abundance does not matter at all. A population of 1 butterfly and 99 birds would have the same diversity as a group of 50 butterflies and 50 birds. An alternative is the Shannon-Wiener entropy, which can be obtained from equation (1) by setting α = 1:

$$. \qquad (2)$$

The Shannon-Wiener entropy satisfies the following useful mathematical properties among continuous functions [Pa10]: (1) Symmetry: the calculated value is independent on how the relative abundance is distributed among the species, e.g. a population of 1 butterfly and 99 birds has the same diversity value as a population of 99 butterflies and 1 bird; (2) Maximum of equality: exists when species exist in equal proportions, i.e. n for all i; (3) Decomposability: the total diversity equals between category diversity plus within category diversity, where categories can be understood as disjunct subsets of species within an ecosystem. While the Shannon-Wiener entropy satisfies these properties, one main limitation of entropy measures is that they do not consider the extent of dissimilarity between objects. Hence, 50 butterflies and 50 birds could have the same diversity as 50 butterflies and 50 beetles.

Distance measures are able to take the extent of dissimilarity between objects into account. For this purpose, distance measures require a formal representation of the characteristic object attributes [Ba06]. Once these characteristic attributes are defined, the diversity is calculated based on some distance function $d(s_i, s_j)$, where s , represent two objects of a set S. In addition to the distance function, one needs to define a mechanism to aggregate the pairwise distance between all objects in S. By simply adding up all pairwise distances, diversity would increase each time an already existing object is added to a set. Although, one should assume that adding a bird to a population of 99 birds and 1 butterfly, should rather decrease than increase diversity. An alternative approach is the recursive aggregation algorithm, developed by Weitzman [We92]: Iteratively increase the diversity value W(S) by adding the distance of an object s_i to the set S, where s_i is the object with the least distance to S. In the beginning S is an empty set and one needs to define a random object to start the recursive aggregation. The main shortcoming of this approach is, that in general the result of the measure is path-dependent and does not provide a unique value. One can show that the outcome of the measure is only uniquely defined, if the pairwise distances are ultrametric, i.e. when for three possible pairwise distances between any three objects, the two greatest distances are equal [Ba06]. This is however, a strong precondition for the assessment of diversity.

Attribute measures offer a more general alternative and were first proposed by Nehring and Puppe [Ne09]. The measure solely accounts for the unique attributes represented by a set of objects and does not consider their relative abundance. This is based on the presumption that the number of distinct attributes accounts better for the diversity of a set than the number of objects. The approach can be applied in different disciplines. The flexibility, however, comes with the cost that one needs to preselect a set of attributes and weights. The predefined attributes of the set Ω are added up by a weighting function:

(3)

where λ represents weights over the different attributes $f \in F$. As stated in [Pa10], depending on whether the weights originate from some objective criterion or subjective criterion, one either has a diversity measure or a value function. The weights should therefore reflect the purpose of the diversity analysis. An additional aspect of the applicability of the multi-attribute approach is the transferability of attributes and the separability of their functions [Pa10]. This depends on the specific assumption over diversity in the respective domain. Just as there are normative judgements needed on the purpose of diversity in biology and in economics, one needs to define what diversity is supposed to represent for the assessment of process variants.

3 Application of diversity measures in BPM

To measure how much diversity is represented by the LCM and the GCD in Fig. 2 one could define each possible execution trace of a process diagram as a single object, similar to a species of an ecosystem. Hence, there is one possible *execution trace* in the LCM and two in the GCD. For the calculation of the *entropy measure* one needs furthermore to define, how often each trace has been executed. For this purpose, following case frequencies will be assumed: <r, h, a> is executed 10 times, <r, c1, a> is executed 9 times and <r, h, c2, a> is executed 1 time. Accordingly, the relative abundance of cases and the Shannon-Wiener entropy for the LCM and the GCD can be calculated. For the calculation of the *distance measure* the Levenshtein distance is used, as described in [Wo06]. Based on the two possible traces in GCD, the outcome of the Levenshtein distance is 2. Since there are only two traces, the value does not change when the Weizman's recursive aggregation algorithm is applied. For the calculation of the *attribute measure* three attributes are assumed to be relevant, which will all be equally weighted by λ=1: number of transitions (events), number of resources and total cost options. For the LCM this results to the simple addition 3 + 3 + 1, for the GCD the outcome equals 5 + 4 + 2. Based on the described premises and calculations, Tab. 1 shows the calculated diversity measures for the LCM and the GCD.

	Traces	**Entropy M.**	**Distance M.**	**Attribute M.**
LCM	1	1	0	7
GCD	2	0.723	2	11

Tab. 1: The 4 diversity measures calculated for the LCM and the GCD in Fig. 2

One can see that only the entropy measure assigns a higher diversity value to the LCM than to the GCD. In the case, where a set only consists of one variant the entropy measure does not have much meaning. The other measures however indicate that the

GCD represents a higher variety of process attributes than the LCM. The most distinctive evaluation is given by the attribute measure, which contains the most information about the two diagrams. This small example demonstrates that existing diversity measures can be applied in BPM. But it also becomes clear that diversity measures need to be carefully selected, in order to provide a meaningful outcome. This also applies for the selection of the considered process attributes.

4 Conclusion

This short paper introduces the idea of applying diversity measures for the improvement of information aggregation in BPM. The approach can be distinguished from other aggregation approaches [Ro17] based on the diversity consideration of selected process attributes. Four common diversity measures have been introduced and tested based on two process diagrams. Several issues, such as the consideration of unfinished traces and process loops have been ignored so far. In addition, the interpretation of the diversity measure outcomes leaves a lot of room for open questions. In order to give a meaningful assessment of business process diversity one should however, consider the diversity in relation to other process aspects, such as performance and flexibility. Only then it is possible to answer whether process diversity can contribute to the improvement of the information aggregation or not. The development of meaningful diversity measures for BPM will hopefully help to answer this question in the future.

References

[Aa02] W. Van der Aalst und T. Basten, „Inheritance of workflows: An approach to tackling problems related to change," Theoretical Computer Science, 270, 1-2, pp. 125-203, 2002.

[Ne09] K. Nehring und C. Puppe, „Diversity," The Handbook of Rational and Social Choice, pp. 298-322, 2009.

[Ba06] S. Baumgärtner, „Measuring the Diversity of What? And for What Purpose? A Conceptual Comparison of Ecological and Economic Biodiversity Indices," SSRN, 2006.

[Pa10] S. E. Page, „Measuring Diversity," in Diversity and Complexity, 2. Hrsg., Princeton University Press, 2010.

[We92] M. Weitzman, „On Diversity," The Quarterly Journal of Economics, Bd. 107. Jg., Nr. 2, pp. 363-405, 1992.

[Wo06] A. Wombacher und M. Rozie, „Evaluation of workflow similarity measures in service discovery," in In Service-Oriented Electronic Commerce, Proceedings zur Konferenz im Rahmen der Wirtschaftsinformatik 2006. Gesellschaft für Informatik e.V., 2006.

[Ro17] M. L. Rosa, W. Van der Aalst, M. Dumas und F.P. Milani, „Business process variability modeling: A survey," ACM Computing Surveys (CSUR), 50(1), pp. 1-45, 2017.