Piloting an Empirical Study on Measures for Workflow Similarity

Andreas Wombacher, Maarten Rozie
University of Twente,
7500 AE Enschede, The Netherlands
a.wombacher@utwente.nl, m.rozie@student.utwente.nl

Abstract

Service discovery of state dependent services has to take workflow aspects into account. To increase the usability of a service discovery, the result list of services should be ordered with regard to the relevance of the services. Means of ordering a list of workflows due to their similarity with regard to a query are missing. This paper presents a pilot of an empirical study on the influence of different measures on workflow similarity. It turns out that, although preliminary, relations between different measures are indicated and that a similarity definition depends on the application scenario in which the service discovery is applied.

1. Introduction

A service oriented architecture is based on services maintained by independent service providers. Services are made available to other parties, such that they may use/invoke these services. Usually two types of services are differentiated: stateless and state dependent services. While stateless services can be invoked by a single request-response communication, state dependent services maintain an internal state and require several request-response communications. In the following, we are focusing on state dependent services interpreting the internal behavior as a workflow.

Once a service has been set up there are different ways of advertising the service. One potential option is publishing the service description at a service repository, which allows parties to search for service descriptions. The result of such a search is a list of service descriptions including endpoint references. A user of such a search facility expects a functionality comparable to other web search engines. In particular this includes that (i) the search results are ordered due to their relevance with regard to the stated query, and (ii) in case there is no exact match of the query the most similar matches should be returned.

Several repositories have been proposed focusing on dif-

ferent aspects of service descriptions like e.g. [1, 17, 16]. Since we are interested in state dependent services, we focus on those engines supporting behavior/workflow descriptions of a service as a search criterion. An example repository has been proposed in [21]. However, this approach provides a list of matches without ordering the results and without providing the required fuzziness in the evaluation of search criteria. Be aware that the required notion of relevance is the relevance of the results to the human user.

One means to address these issues is to use a notion of workflow similarity as a relevance measure like e.g. in [22, 6]. In particular, the more similar the query is with regard to a behavior/workflow describing the service the more important it is for this query. However, such technical similarity measures have to be evaluated with regard to their applicability, i.e., whether the derived order of search results really represents the relevance of the services as expected by a human user.

Therefore, the aim of the empirical study is to explore the intuitive understanding of workflow similarity by humans. In particular, we are aiming to understand which factors of a workflow specification are most important to workflow similarity, i.e. - applied to the service repository mentioned before - representing the relevance of a service. To get a good understanding, a questionnaire is designed based on hypothesis on the importance of workflow specification aspects. The questionnaire has been answered by several people as a pilot study of an empirical study showing first tendencies and giving indications on how the questionnaire can be improved. The pilot indicates already that there are several workflow specification aspects with different relevance although their importance with regard to the similarity definition depends on the stage in the service life cycle where the service discovery is applied. Further, it turned out that the aggregation of the questionnaire results is difficult because human interpretation has been required. Therefore, the design of the questionnaire has to be adapted for the run of the empirical study to support an automatic aggregation of results.

In Section 2 the design of the research is discussed by

introducing the formal workflow model, the set of hypotheses, and the design of the questionnaire. Section 3 presents the analysis of the conducted pilot study, an interpretation of the results, and lessons learned. The paper concludes with related work (Section 4) and conclusions (Section 5).

2. Research design

Since no agreed technical definition of workflow similarity exists, we investigate the human intuitive understanding by performing an empirical study. The goal of this study is to get an understanding of the different aspects of a workflow description on the similarity measure. Potentially, each individual will have a different intuition of what is important for the similarity of workflows. Therefore, the best way to conduct an empirical study is to ask multiple persons, thus, the results will be more reliable.

The study will be based on a group of technical workflow specialists in a first run. This is because the involvement of domain specific business process respondents requires a discussion of more complex example workflows since semantics of the workflows is expected to play a more important role. We are planning to extend this study to this group of respondents in future work.

Further, we decided to use a questionnaire as a means to ask multiple persons, because it is efficient with regard to costs and time. The questionnaire can easily be sent via email to international participants without having travel expenses. Other means to conduct this empirical study could have been interviews. However, we were concerned about the comparability of the results of such interviews¹.

For this empirical study we base the questions on a simple workflow model. As a basis for designing the questions, several hypotheses are used representing the relation between different workflow description aspects. Based on these hypotheses and by using the simple workflow model the questionnaire can be set up [7].

2.1. Formal workflow model

The questionnaire uses Finite State Automata [8] as the simplest possible model to represent workflows. We decided not to use more complex models at a first glance because the more expressive the workflow model the harder conclusions can be derived at the end. However, the results derived from this study are applicable to the service discovery scenario since the before mentioned repository [21] is based on slightly extended Finite State Automata [20].

A Finite State Automaton is based on states represented as circles, a start state, a set of finite or accepting states

¹Probably a study involving business people may require interviews supporting human interaction to avoid misunderstandings since the business scenarios are most likely more complex.

represented by circles with thick lines, and labeled transitions represented as directed arcs. In particular, a labeled transition means that a state is changed when a certain message is either sent or received. An automaton describes the potential execution sequences of a workflow which is also called the language of an automaton. Example Finite State Automata are depicted in Figure 1.

2.2. Hypotheses

We expect that several aspects of a workflow influence the workflow similarity. In particular, the aspects are related to language, structure and semantics, where each can more easily be represented by a single measure. With language we mean the possible execution sequences of an automaton. Structure means the structural representation of an automaton comparable to a directed graph. With semantics we mean the semantics of the used transition labels determining the semantics of the complete workflow. In the following we are using the measures to discuss the influence of each aspect of a workflow description on the similarity measure for brevity.

Since there is no clear understanding on how the different measures depend on each other we come up with a set of hypotheses. It is quite obvious that semantics is an important measure for similarity but on the other hand side it is quite hard to evaluate its impact using questions. This is because changes in semantics are always effecting the remaining two technical measures too. Therefore, the semantic measure is explicated in a single hypothesis only:

H0 : Semantics plays an important role This hypothesis is considered implicitly in all questions of the questionnaire by using semantically meaningful workflows. In particular, we use RosettaNet Partner Interface Processes (or PIPs) [14] as transition labels. Examples of the used labels and brief descriptions of their semantics are given in Table 1. Be aware, that some of the PIPs are covering two messages which are usually request and response messages. We use the notation of labels without prime (like e.g. *a* in PIP 3A2) for request messages and labels with prime (like e.g. *a'* in PIP 3A2) for the corresponding response message².

The example automaton depicted on the left hand side of Figure 1 uses the labels described in Table 1. This workflow starts with a request for a purchase order (transition labeled *p*), followed by an acceptance of the purchase order (*p'*). Then, either an invoice for this specific order (*i*) or an invoice covering a certain time span (*b*) is sent. The customer can now choose to pay the order (*r*), after which the order is shipped (*n*), or to send a cancellation request (*c*) followed by an cancellation confirmation (*c'*).

²During this study we assume that a response will always have a positive return message, thus we are not explicating exception handling.

Table 1. Examples of used PIPs

Code	Label	Name
PIP3A4	p and p'	Request purchase order (PO)
PIP3A9	c and c'	Request PO cancellation
PIP3B2	n	Notify of advance shipment
PIP3C3	i	Notify of invoice
PIP3C5	b	Notify of billing statement
PIP3C6	r	Notify of remittance advice

The structural and language based measures are easier to handle compared with the semantics one. Thus, we define a set of hypotheses, where a hypothesis states how two technical measures are related to each other with regard to their importance. In particular, a hypothesis is stated as a comparison of two measures. The hypotheses cover different levels of granularity, thus, one hypothesis can be subsumed by another. This is because there might be a need to differentiate more specific technical measures. The expectation is that the results for hypotheses addressing more specific measure will be less explicit. Here follows a short description of the additional hypotheses.

H1: Similar language > Similar structure The hypothesis is of high granularity and states that the language of an automaton is more important than its structure. An automaton with similar language will be considered more equal than an automaton that has a similar structure.

H2: Similar transition labels > Similar sequence The hypothesis states that the transition labels used by an automaton are more important than the actual order of these labels. Thus, automata that have a similar set of labels in a different sequence are considered to be more similar than automata that have a similar sequence, but contain some different messages.

H3: Few important transitions similar > Lot less important transitions similar The hypothesis is about the difference in importance of transitions. It assumes that having a few important transition labels in common will make an automaton more similar than having a lot of less important labels in common.

H4: Being a super-automaton > Similar structure

H5: Being a sub-automaton > Similar structure Hypotheses 4 and 5 are more specific than hypothesis 1. In particular, in H1 language similarity is based on the number of commonalities of the languages (intersection), in these hypotheses language similarity is based on subsumption of languages. In particular, hypothesis 4 expects additional

sequences to be provided by the automaton in the repository, while hypothesis 5 expects the additional sequences in the query automaton. In both hypotheses the corresponding language measure is more important than the structural measure.

H6: Having extra transitions in same path > Having extra paths Hypothesis 6 is stated as a structural measure only, which influences also language measures. Having extra transitions in the same path will not change the general structure of an automaton. Further, having extra paths will change the structure but not necessarily the original language. In fact, the automaton with extra paths are super-automata. Thus, this hypothesis conflicts with H4 because structure measure is considered more important than some kind of language measures.

H7: Having transition as a loop > Not having transition

H8: Having different transition > Not having transition Hypotheses 7 and 8 are addressing the impact of changing a single transition of an automaton. Hypothesis 7 can be viewed as a specific case of hypothesis 4 or 5. In particular, a transition in one automaton is represented as a loop in another automaton. Thus, the latter one subsumes the language of the former one. Further, this modification of the second automaton does not affect the structure of the first automaton too much.

Hypothesis 8 is another refinement of hypothesis 1. The hypothesis varies automata by keeping the structure of an automaton intact but relabeling a transition by another label not yet contained in the alphabet of the original automaton. As a consequence, the structure is not changed, but the part of the language influenced by the changed transition is disjoint.

H9: Having a path in common > Similar structure The final hypothesis considers whether having a single path in common is more important than having the same structure. While hypothesis 1 compares complete languages, in H9 the intersection of languages is applied.

2.3. Questionnaire

Based on the above hypotheses the questions of the questionnaire are constructed. As a basis for the empirical study it is necessary that we get a reasonable amount of filled-in questionnaires back each providing a high data quality. Thus, we are supposed to provide a questionnaire which can be answered in a reasonable amount of time and we have to add redundancy in the questionnaire to detect random answers. To reduce the complexity of the questionnaire we keep the number of questions as low as possible and limit the number of different automata used in the questionnaire.

Figure 1. Question 1

However, to achieve good data quality each hypothesis is at least tested in two different questions.

Each question is structured in the following way. A reference automaton and a set of either three or four solution automata (A, B, C, D) are given. A respondent has to order the solution automata by similarity with respect to the reference automaton. If a respondent finds multiple solution automata equally similar to the reference automaton, he can assign several automata to the same position of the order. Respondents are also asked to state their reason on how they derived the provided order. An example question is depicted in Figure 1.

The solution automata for each question should only differ on measures relevant to the hypothesis for which the question is used. So the given answers will indicate the correctness of the hypothesis and as a consequence the importance of the corresponding measures addressed in the hypothesis.

The requirement that the automata are semantically meaningful and the limited set of used PIPs restricts the potential number of automata. We use the provided reasons of the respondent to distinguish whether the ordering of solution automata is due to the corresponding hypothesis or due to semantic similarity. The questionnaire is available at [15]. On the two left hand side columns in Table 2 the different questions of the questionnaire are related to the corresponding hypothesis. Be aware that there has been no specific question for hypothesis 0 (semantics) since the influence of semantics will be unavoidable in each of the questions.

After the design of the questionnaire we tested whether the questionnaire and its description are understandable and whether the time to fill out the questionnaire is reasonable. Thus, a project independent person filled out the questionnaire, which took him 60 minutes. Further, he provided valuable feedback on the descriptions contained in the questionnaire which resulted in some modifications. The time of 60 minutes seemed to be reasonable, thus we added this information to the questionnaire as well.

Figure 2. Statistics of consistent answers

3. Analysis

In the following, the results of the questionnaire are discussed. First, the data collection is described followed by an evaluation of the results with regard to each hypothesis. Then a summary on the influence of the different measures on the workflow similarity is presented including the lessons learned.

3.1. Data Collection

For piloting the study, the questionnaire has been sent to 27 international respondents. Although most of them are working in the workflow domain, they have different backgrounds and different areas of expertise, like e.g. inter-organizational workflows, workflow matchmaking, or semantic service composition.

We received 12 responses from Germany, Austria, Italy, UK, Canada, US, and the Netherlands. In most cases the respondents made use of the reason field in the questionnaire. The questions with regard to the different hypothesis have been answered more or less consistently. The highest percentage of consistent answers over all questions related to a hypothesis has been 83% of the respondents while the lowest percentage has been 67% of the respondents. The distribution of the percentage values per hypothesis is depicted on the left hand side of Figure 2. The lowest percentage of consistent answers per respondent over all hypotheses is 44% while the highest is 100%. The distribution of the percentage values per respondent is depicted on the right hand side of Figure 2.

It turned out that it was impossible to derive results from question Q10 and Q15. In case of Q10 the results provided did not reflect the corresponding hypothesis. This is because the only automaton in the solution set that was not a super-automaton of the reference is an automaton that had no real semantic meaning. Also in Q15 the semantic measure overruled the corresponding hypothesis. Thus, these questions have to be redesigned for the study.

Table 2. Summary of results per question

Hypotheses	Question	support	max	oppose
H1	1	7	7	1
	5	5	3	2
H2	7	2	2	6
	8	4	2	5
	18	2	2	5
	21	5	2	6
H3	9	3	1	-
	22	3	2	-
H4	3	8	3	0
	10	-	-	-
	13	9	4	0
H5	6	9	4	1
	17	7	3	0
H6	11	5	2	6
	14	9	5	3
H7	2	6	5	5
	16	6	6	2
	23	5	4	3
H8	12	4	1	7
	19	4	1	5
H9	4	3	3	2
	15	-	-	-
	20	6	3	6

3.2. Hypotheses

The results of the hypotheses will be checked by looking at the number of respondents supporting or opposing a hypothesis taking into account the provided reason (see also the corresponding columns in Table 2³). In particular, a hypothesis will be accepted if the number of supporting respondents is high and the provided reasons are related to the hypothesis. For example, in case a large number of opposing answers have a reason that is not mentioned in the hypothesis this will be counted less severe for interpreting the results as opposing answers with reasons opposing the hypothesis. However, questions with reasons unrelated to the hypothesis have to be considered for change in the study. In the discussion of the results sometimes a question is referenced by its number within the questionnaire.

H1: Similar language > Similar structure In the first question (Q1) of this hypothesis the number of support depends on the rate of similarity of the language. One of the language similar automata supported the reference language completely while the other missed one transition. Eleven of the twelve respondents placed the complete language similar automaton on the first place, seven of them also put the partial language similar automaton in front of the struc-

tural similar automata. The four respondents considering the structural similar automaton on position two argued on the semantic similarity to the reference automaton. The one respondent opposing the hypothesis gave the structure of the automata as a reason.

In the second question (Q5) both language similar automata are not completely equal to the reference. Therefore, the support is less in this question. Ten respondents placed one of them in first place, while five put both in front of the structural similar automata. Again semantic similarity influenced the results by rating one structurally different automaton better due to its semantic similarity (five respondents). The two respondents disagreeing with the hypothesis gave as reason the structure, namely the number of end-states.

Overall there is a lot of support for this hypothesis, especially if the language of the solution automata are exactly the same as the reference automaton.

H2: Similar alphabet > Similar sequence The respondents disagree on this hypothesis. Two groups of respondents can be clustered based on the results of the corresponding questions. Two respondents agree with the hypothesis for all questions, while five respondents always disagree. Disagreement comes from respondents looking at the supported behavior of the automata, which results in a last place for automata with similar labels but slightly different sequence.

Semantic similarity of messages b and i has an impact again. In some answers the order of labels is considered less important than the similarity of b and i , while in other questions this is reversed.

The conclusion of the results is that the level of support for this hypothesis depends on how much a sequence is changed in an automaton with similar labels. Only having similar labels as the reference automaton is not sufficient, since the remaining sequences can not differ too much, otherwise the number of supporting answers decreases.

H3: Few important transitions similar > Lot less important transitions similar A lot of different answers are given to the questions for this hypothesis. Some of the respondents focus on the supported behavior for ordering the solution automata. Others use the importance of messages to derive an order, like e.g. considering an acknowledgment message c' less important than corresponding cancellation message c . Finally, some prefer to remove a complete request-response pair instead of just the response message. As a consequence the hypothesis can neither be confirmed nor inverted, but there is an indication that it could be ignored for setting up a similarity measure.

³The table contains in addition a column *max* indicating the maximum number of respondents providing exactly the same order.

H4: Being a super-automaton > Similar structure

This hypothesis is supported by eight respondents on all questions. The respondents that support the hypothesis consider the behavior of the solution automata. When the reference automaton is a sub-automaton of a solution automaton, then the behavior of the reference automaton is automatically supported by the solution automaton. So when looking at the behavior, the super-automaton will be placed on the first position.

The respondents opposing the hypothesis used different reasons to disagree with the hypothesis. One of them based the answer on the similarity of messages *b* and *i*. The other found the additional behavior of the super-automaton worse than the behavior of the structural similar automaton.

The high support for this hypothesis is consistent with the acceptance of H1. We expected this due to the close relation of H1 and H4.

H5: Being a sub-automaton > Similar structure Seven respondents support the hypothesis in both questions. They focus on the supported behavior of the solution automata. Because a sub-automaton of the reference automaton has most paths in common, this automaton is placed at the first position.

Again the semantics influences the result. The similarity of message *b* and *i* overlaps the hypothesis under investigation. There is only one respondent giving a reason opposing the hypothesis by preferring the structure similarity over sub-automaton based language similarity.

Being a sub-automaton of the reference automaton seems to influence similarity. Disagreement for this hypothesis is based on the semantics of the workflows rather than the structure of the workflows. Therefore, the corresponding questions must be changed for the study.

H6: Having extra transitions in same path > Having extra paths

Adding extra transitions as a postfix of a path is preferred by a large number of respondents. However, adding transitions in other parts of a path is much less supported. The two related questions reflect this difference.

The first question of this hypothesis (Q11) can be seen as a sort of inverse question to hypothesis 1. In this question two solution automata are constructed by adding extra transitions in the same path as the reference automaton, which keeps the structure intact. The other solution automata are super-automata of the reference automaton supporting its language, but adding transitions and states.

The respondents are split on this question. Six found it more important that an automaton supports the behavior of the reference automaton and did not mind the false positives given by the additional behavior of the super-automata. Five preferred the automata with additional transitions in the path, two mentioning the structure as a reason for this

choice. The last respondent placed one of the extra transition automata in first place and the other in last place. The reason for this ordering has been the meaning of the complete workflow.

In the other question (Q14) the additional transitions are placed as a postfix to the paths of the reference automaton. A majority of the respondents found that adding transitions this way does not have as much impact on similarity of automata as adding complete new behavior. Nine respondents agreed with the projected order of the hypothesis, three disagreed.

As a summary, this hypothesis has limited support which is mainly dependent on where the messages are added to the automaton.

H7: Having a transition as a loop > Not having the transition

Five respondents support the hypothesis. The major reason for preferring automata with the transition as a loop is that those automata can support the language of the reference automaton. The extra paths created by the loop are considered less important for the similarity.

Two respondents disagree completely, they find the change in behavior by adding a loop too significant and prefer removing the transition.

Other respondents provided inconsistent answers. One of them ordered the solution automata by the number of common paths with the reference automaton. Another one considered the loops introduced in one question more significant on the possible behavior of the automaton than in the other questions. Thus, this hypothesis has a weak support.

H8: Having a different transition > Not having a transition

The respondents can be divided into two groups for this hypothesis. There are four supporters of the hypothesis and seven respondents who disagree. Reasons for supporting the hypothesis are the structure of the workflow and keeping the sequence of other transitions unchanged. The most often mentioned reason for disagreement has been that it is better to remove a transition than replacing it with another one.

Overall the result is that adding and removing a transition can be considered equally important for similarity as changing a transition. This view is in accordance to the edit-distance of strings where adding and deleting a transition has the same cost as changing a character.

H9: Having a path in common > Similar structure

Three respondents looked at the behavior of the automata, preferring the solution automata that has the most common paths with the reference automaton. This method resulted in answers supporting this hypothesis.

There have been also respondents who opposed the hypothesis where the main reason has been the similar meaning of messages b and i . No one mentioned structural elements as a reason for preferring a different order. However, the strong influence of the semantics of the message limits the conclusions that can be derived from this hypothesis.

H0: Semantics plays an important role As stated in section 2.2, semantics on similarity is influencing every question. Which has been one of the results of the discussion of the above hypotheses.

The questionnaire contained three questions with two automata in the solution set that both differed from the reference automaton on one message. One of them had a semantically similar message, the other a semantically different message. Eight of the twelve respondents placed the automaton with the semantically similar message before the other in all questions. The four other respondents rate the automata equally similar because they looked at the possible sequences of the automata.

From the results of H2 can be concluded that the influence of semantics becomes greater when the sequence of the solution automata is changed more drastically. When automata have two messages switched, only two respondents prefer a semantically similar automaton. If automata have three messages ordered differently, six respondents choose a semantically similar automaton. We observed that some respondents provide different orders for comparable questions while mentioning the same reason. This makes clear that the influence of semantics is quite different.

When looking at the results of the questions related to H4 the influence of semantics is low. When there are super-automata among the solution automata the semantically similar automaton is only preferred in three of the twenty-four answers. For H5 the influence of semantics is greater. Five respondents prefer the semantically similar automaton for both questions, while five respondents prefer the sub-automata.

3.3. Summary

As a summary there are three hypotheses, where language is compared to structure (H1,H4,H5), which have a lot of support. As a consequence the language measure is more important than the structural measure for workflow similarity. Even if the language is not exactly the same, but smaller or greater, the automata with the same structure are considered less similar.

Super-automata are also considered more similar than automata with extra transitions before or within the paths of the reference automaton, as can be concluded from H6. However, when the extra transitions are added as a postfix to the automata, those are considered more similar.

H2 indicates that a similar alphabet is important, but when the structure and sequences are changed too much, automata with different alphabets but more similar sequences are preferred.

From H7 and H8 is indicated that replacing a single transitions by either a loop or another transition and removing the transition have the same impact on similarity.

The remaining two hypotheses (H3,H9) were unusable. For H3 the importance of individual transitions is too dependent on the specific situation. In H9 agreement differs a lot for each question, disagreement is caused by the semantics of the workflows.

The influence of semantics (H0) is low when there are super-automata or automata with a similar language in the solution automata. These automata are preferred over semantically similar automata. When solution automata differ with regard to language, the semantically similar automata are preferred. Thus, there are cases where semantic measures are significant for the workflow similarity although the concrete influence in general remains still unclear.

3.4. Interpretation of preliminary results

The result of this pilot is a better understanding of the measures influencing workflow similarity. However, we do not feel comfortable to come up with a similarity definition specifying the influence of the different measures exactly. This is because the impact of different measures depends on the automata used. Therefore, we are convinced that a technical similarity measure will at most be able to approximate the results derived in this empirical study.

Further, the inconsistencies in answering questions, which is up to 33% at a maximum for certain hypotheses, indicates that the mind set of the respondent is also a relevant factor influencing the definition of the similarity measure. This aspect has not been investigated in this pilot. In particular, we are convinced that the application domain of the workflow similarity has an impact on its definition, that is, the importance of the different measures.

In the following we illustrate the influence of the application scenario on the similarity measure using the different occurrences of service discovery in a Web Service life-cycle. In particular, we indicate the most relevant measure for each of the relevant phases. The life-cycle model consists of the following phases [2]: business process improvement, service interface design, service design, service construction, service testing, use/reuse service, deployment, operations and measure results. Service discovery may be applied in the design, the construction, and the operation phase.

Design

Service design is the phase where the business must decide which services to create for delivering the business require-

ments. The most important attribute in this phase is semantics, because the design of the new service can still be changed in this phase. Semantically similar services can be used to get ideas how to design the new service. In this case semantic similarity is used on a workflow level. The language attribute can also be useful in this phase. Language similarity measures can be used to see if they provide additional transitions that are not considered yet. A structural measure is not very useful, however, searching for structural similar services can be used to see if the granularity of the defined services is in accordance with other services.

Construction

In the service construction phase services are created, either by programming them or by making a composition of existing services. The service developer wants to reuse (parts of) existing services as much as possible. Services that are suitable for reuse are services that have complete paths in common with the design. Therefore, the most important attribute for similarity in this phase is the language of services. Semantics also has an influence on similarity here. When the service developer does not find language similar services, but there exist semantically similar services, he can contact the designer if these can be used for a redesign. Structural similar services cannot be used in this phase.

Operation

During the operation phase the created services are used to support business processes. Services must be able to work with each other. The compatibility depends on transitions and sequences of messages exchanged. Therefore, the most important measure is language. The most useful services will be a service that is equal or a super-automaton, which means both options support the same behavior as required. Alternatively a sub-automaton can provide a part of the required behavior. Such a service can be used as a basis for service composition. During usage, semantic similarity of services is less important, because services are not capable of translating semantically similar messages to the message they require. Services with a similar structure are not applicable during the operation phase.

3.5. Lessons Learned

Due to the above described results of the pilot, several questions have to be considered for redesign to improve the presentation of the hypotheses. By changing questions, also the aggregation of results can be improved, which must be possible without analyzing the reason field. Further, hypotheses on semantics have to be stated explicitly relating semantics and language measures. Further, in the introduction of the questionnaire we have to describe the intended application scenario addressing the operation phase of the service life-cycle. However, we are satisfied with the indications we got from the pilot and the data quality we could

achieve already.

4. Related Work

We are not aware of any empirical evaluation of similarity of workflows. However, there has been a lot of work on standardizing messages (like e.g. EDIFACT), process parts (like e.g. RosettaNet PIPs [14]), and processes (like e.g. IOTP [3]) for specific use cases in specific domains. The design of these frameworks requires an implicit similarity notion of message structures and processes to decide e.g. which process parts are being standardized while others are considered to be not significant or different enough.

The bundling of functionality into clusters or components is a topic related to software engineering. In particular metrics have been developed to indicate the quality of an actual cluster. In [13] such a metric has been applied on designing workflows in finding the right level of abstraction of tasks within an acyclic workflow model based on data flow. However, such a metric specifies the quality of a workflow and does not contribute to compare two workflows.

With regard to measures of workflow similarity (in the sense of a closeness measure of workflows) different scientific research areas have to be considered⁴. One dimension which is really hard to address is the behavioral aspects covered in all kinds of semantic service descriptions like e.g. WSMO, WSDL-S, or OWL-S. Dependent on the underlying models different operations and reasoning facilities are available. However, the elicitation of semantics from existing data is quite a challenge and ongoing research.

There exist different distance measures for strings like for example the Hamming distance used in information theory [18] or the edit distance usually applied on strings in the context of text. The edit distance (or Levenshtein distance) [9] between two strings is the smallest number of substitutions, insertions, and deletions of symbols that can be used to transform one string into another. This definition based on a single string can be extended quite easily to a set of strings, i.e. languages. However, this extension does not work in case at least one language is infinite. In [12] an approach where costs are assigned to each change operation. Actually, the approach calculates the minimal distance of a string accepted by the first automaton with a string accepted by the second automaton. The issue with this approach is that the similarity drills down to the difference of two strings, which is quite unspecific in case the languages contain a lot of strings.

The issue with infinite languages can be addressed by using an abstraction of the infinite language to a finite representation [22] e.g. based on an ngram representation of

⁴A detailed discussion of the different approaches and possible measures is contained in [22].

automata [10]. The draw back of any abstraction is the loss of information resulting in introducing false positives.

An automaton can be interpreted as a directed graph and therefore graph similarity measures can be applied. An exemplary similarity measure based on edit distance has been proposed in [4] addressing graph isomorphism, while [11] addresses subgraph isomorphism. Another structural approach for reconciliation of processes is presented in [5] providing a similarity measure. The approach focus on the common alphabet of two automata and removes the exclusively used messages of the alphabets. Further, automata transformation rules, as specified in [19] for Workflow-Nets, are used to transform both automata to the same automaton using only the shared alphabet. In either case, the language aspects are neglected and only structural aspects are considered, which is not sufficient

5. Conclusion and Future Work

In this paper, the need of a workflow similarity definition is motivated in the Web Service domain and the impact of several similarity measures on this definition has been indicated by piloting the empirical study. It seems that the impact of the similarity measures on a concrete similarity definition depends on the application scenario in which such a definition is used. However, the results of the pilot indicate that several hypotheses on the general relation of several measures are a good starting point. Further, it turned out that some questions have to be revised and questions addressing the particular influence of semantics should be added.

Future work will address more complex workflows involving additional respondent groups, in particular from the business side. Further, we will develop similarity definitions for service discovery in the design and the operational phase.

Acknowledgment

Thanks to Warren Blanchet, Ralph Bobrick, Carine Courbis, Schahram Dustdar, Pascal van Eck, Bendick Mahleko, Marco Pistore, Manfred Reichert, Stefanie Rinderle, Jianwen Su, Barbara Weber, and Zlatko Zlatev for filling out the questionnaire.

References

- [1] A. S. Ali, O. F. Rana, R. Al-Ali, and D. W. Walker. UD-DIe: An extended registry for web services. In *Proc. of the 2003 Symposium on Applications and the Internet Workshops (SAINT-w)*. IEEE Computer Society, 2003.
- [2] M. Brodie. Opportunities for semantics in service-oriented architectures. *Web Service Semantics Workshop*, 5 2005.
- [3] D. Burdett. Internet open trading protocol - IOTP - version 1.0. <http://www.ietf.org/rfc/rfc2801.txt>, 2000.
- [4] G. Chartrand, G. Kubicki, and M. Schultz. Graph similarity and distance in graphs. *Aequationes Mathematicae*, 55:129–145, 1998.
- [5] Z. Du, J. Huai, Y. Liu, C. Hu, and L. Lei. IPR: Automated interaction process reconciliation. In *Proc. of IEEE/ACM Int'l. Conf. on Web Intelligence (WI)*, 2005.
- [6] N. E. Fenton and S. L. Pfleeger. *Software Metrics - A Rigorous and Practical Approach*. 1996.
- [7] A. Fink and J. Kosecoff. *How to Conduct Surveys, A Step-by-Step Guide*. SAGE Publications, 1985.
- [8] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, 2001.
- [9] L. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals,. *Soviet Physics-Doklady*, 10(8):707–710, 1966.
- [10] B. Mahleko, A. Wombacher, and P. Fankhauser. Process-annotated service discovery facilitated by an n-gram based index. In *Proc. of IEEE Int'l. Conference on e-Technology, e-Commerce and e-Service (EEE)*, pages 2–8, 2005.
- [11] B. T. Messmer and H. Bunke. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):493–504, 1998.
- [12] M. Mohri. Edit-distance of weighted automata: General definitions and algorithms. *Int'l. Journal of Foundations of Computer Science*, 14(6):957–982, 2003.
- [13] H. A. Reijers and I. T. P. Vanderfeesten. Cohesion and coupling metrics for workflow process design. In *Proc. Int'l. Conf.: Business Process Management (BPM)*, LNCS 3080, pages 290–305, 2004.
- [14] RosettaNet. <http://www.rosettanet.org>, 2004.
- [15] M. Rozie and A. Wombacher. Questionnaire of the empirical workflow similarity study. http://wwwhome.cs.utwente.nl/~wombachera/papers/questionnaire_v1.0.zip, 2005.
- [16] N. Srinivasan, M. Paolucci, and K. Sycara. Adding OWL-S to UDDI, implementation and throughput. In *First Int'l. Workshop on Semantic Web Services and Web Process Composition (SWSWPC)*, 2004.
- [17] K. Sycara, M. Paolucci, A. Ankolekar, and N. Srinivasan. Automated discovery, interaction and composition of semantic web services. *Journal of Web Semantics*, 1(1):27–46, Sep 2003.
- [18] H. Tzschach and G. Hasslinger. *Codes fuer den stoerungssicheren Datentransfer*. Oldenburg Verlag, 1993.
- [19] W. M. P. van der Aalst and T. Basten. Inheritance of workflows: an approach to tackling problems related to change. *Theor. Comput. Sci.*, 270(1-2):125–203.
- [20] A. Wombacher, P. Fankhauser, B. Mahleko, and E. Neuhold. Matchmaking for business processes based on choreographies. In *Proc. of Int'l. Conf. on e-Technology, e-Commerce and e-Service (EEE-04)*. IEEE Computer Society, 2004.
- [21] A. Wombacher, B. Mahleko, and E. Neuhold. IPSI-PF: a business process matchmaking engine based on annotated finite state automata. *Journal on Information Systems and E-Business Management*, 3(2):127–150, 2005.
- [22] A. Wombacher and M. Rozie. Evaluation of workflow similarity measures in service discovery. In *Tagungsband der Multikonferenz Wirtschaftsinformatik - Service Oriented E-Commerce Track (MKWI)*, 2006.