Institute for Software Research

University of California, Irvine

An Empirical Study of Scenario Similarity Measures

Thomas A. Alspaugh
University of California, Irvine
alspaugh@ics.uci.edu

Annie I. Antón
North Carolina State University
anton@csc.ncsu.edu

Laura J. Davis
North Carolina State University
laura@bodeonline.org

September 2003

ISR Technical Report # UCI-ISR-03-7

Institute for Software Research
ICS2 210
University of California, Irvine
Irvine, CA 92697-3425
www.isr.uci.edu

An Empirical Study of Scenario Similarity Measures

Thomas A. Alspaugh
Institute for Software Research
University of California, Irvine
Irvine, CA 92697-3425 U.S.A.
alspaugh@ics.uci.edu

Annie I. Antón
College of Engineering
North Carolina State University
Raleigh, NC 27695-8207 U.S.A.
aianton@eos.ncsu.edu

Laura J. Davis
College of Engineering
North Carolina State University
Raleigh, NC 27695-7534, USA
Laura@Bodeonline.org

ISR Technical Report #UCI-ISR-03-7
September 2003

Abstract: Syntactic similarity measures have been proposed as a technique to support scenario management and provide process guidance in scenario-based requirements analysis. Similarity measures support locating duplication and near-duplication between scenarios, searching in a collection of scenarios, identifying episodes shared among scenarios, and determining dependencies between scenarios. The effectiveness of this technique depends in part on how well syntactic similarity tracks semantic similarity as judged by human analysts. We present ae study that validates syntactic similarity measures using scenarios from the Enhanced Messaging System specification.

An Empirical Study of Scenario Similarity Measures

Thomas A. Alspaugh
Institute for Software Research
University of California, Irvine
Irvine, CA 92697-3425 U.S.A.
alspaugh@ics.uci.edu

Annie I. Antón
College of Engineering
North Carolina State University
Raleigh, NC 27695-8207 U.S.A.
aianton@eos.ncsu.edu

Laura J. Davis
College of Engineering
North Carolina State University
Raleigh, NC 27695-7534, USA
Laura@Bodeonline.org

ISR Technical Report #UCI-ISR-03-7
September 2003

Abstract

Syntactic similarity measures have been proposed as a technique to support scenario management and provide process guidance in scenario-based requirements analysis. Similarity measures support locating duplication and near-duplication between scenarios, searching in a collection of scenarios, identifying episodes shared among scenarios, and determining dependencies between scenarios. The effectiveness of this technique depends in part on how well syntactic similarity tracks semantic similarity as judged by human analysts. We present ae study that validates syntactic similarity measures using scenarios from the Enhanced Messaging System specification.

1 Introduction

The use of scenarios in software development has become increasingly common [14]. Scenarios have certain definite advantages as a means of describing software behavior. Their narrative structure takes advantage of the natural human skills in telling and understanding stories, and their informality makes them more accessible and appealing to those system stakeholders whose technical and mathematical backgrounds are not strong. The use of scenarios for requirements or as a behavioral specification involves a number of challenges, however, especially with the collections of 50 to 500 scenarios that are frequently needed to describe real software systems. When the number of scenarios exceeds what one person can keep in his or her head at once, the problems of scenario management become prominent. Scenario management addresses problems that become markedly less tractable as the number of scenarios for a system increases, such as determining whether a collection of scenarios is complete; determining whether the scenarios are consistent, and managing the inconsistencies between them; detecting and eliminating duplicate and near-duplicate scenarios; organizing and classifying the scenarios for a system; finding the relationships among scenarios, and tracing dependencies among scenarios and between scenarios and other artifacts; process guidance; and scenario evolution.

In our previous work we proposed a strategy that allows software tool support for collections of scenarios and scenario management [1, 2, 4]. Among the components of this strategy are syntactic similarity measures, supported by a representation of scenarios as sets of attribute-value pairs, the use of glossaries (one per attribute) to organize the range of values for each attribute, and a software tool that implements all of these. A syntactic similarity measure indicates similarity of syntactic form, without consideration of semantics (except as it is expressed in syntax) or use of external semantic structures. The specific similarity measures we describe compare two scenarios by comparing the relative number of identical *vs.* non-identical attribute values for the two scenarios; we review the similarity measures in detail below. A syntactic similarity measure is of interest because

it can be computed quickly and automatically. It is useful to the extent that it indicates the degree of similarity that a human analyst would find after a careful examination of the scenarios. As nearly as possible, the measure must identify as similar those scenarios that an analyst would deem similar, while bringing in as few as possible "false positives" that an analyst would not consider to be similar. Ideally, a similarity measure would approximate the degree of similarity that an analyst would find, so that where an analyst would judge one pair of scenarios to be more similar than another pair, the similarity measure would assign a higher similarity to the first pair as well.

Similarity is of pragmatic value in dealing with some of the scenario management problems that arise for a large collection of scenarios: for searching for a particular scenario, identifying unintended duplication and near-duplication, and as a basis for identifying potential episodes (intentionally shared subsequences of events) and other dependency relationships between scenarios [1, 4].

The work presented here was motivated by the ongoing development of a software tool to support scenario-based specification, which had reached the point at which implementation of a similarity measure was possible. This tool, SMaRT (Scenario Management and Requirements Tool), supports analysts as they enter, manage, view, analyze, and work with scenarios and associated episodes, requirements, goals, and conditions [1, 3]. The study presented below provided an evaluation of whether development resources should be used to add our syntactic similarity measures to SMaRT.

In this paper, we present an study in which the results of a family of similarity measures are compared against the judgment of one of the authors (Davis) acting an analyst. The analyst divided a collection of twelve scenarios into disjoint groups of scenarios based on intuitive similarity. The scenarios in each group were judged to be similar to each other, and dissimilar to all the other scenarios. One scenario deemed dissimilar to all others formed a singleton group. At the time this division was done, the analyst was not familiar with the similarity measures that were to be used in order to avoid bias in their favor as much as possible. The effect of this grouping was to mark each pair of scenarios as either similar or dissimilar. No attempt was made to order the 132 pairs in a range from most intuitively similar to least intuitively similar, for reasons of time; creation of such a ranking would form the basis for a future study. The scenarios were selected from those of the Enhanced Messaging System specification [2] by another of the authors (Alspaugh). The goal of the selection was to ensure the presence of pairs of scenarios ranging from quite similar to quite dissimilar.

2 Related work

Syntactic similarity is used as a component of Integrated Scenario Analysis (ISA) [1, 4]. ISA comprises several approaches and techniques:

- representing scenarios as event sequences plus attribute-value pairs,
- using sequence, iteration, and alternation to effectively express event sequences,
- using glossaries to define the values of attributes,
- using glossaries of words and phrases that have system-specific meanings,
- expressing dependency among scenarios with the use of episodes,
- visualizing the dependencies among scenarios by means of the "includes" hierarchy of scenarios and episodes, and
- using syntactic similarity to measure scenario similarity, search for specific scenarios, and identify potential duplication.

These approaches mutually reinforce each other and are amenable to automated support. Automated support for ISA is provided by SMaRT. ISA provides a foundation for scenario management and leads to scenario collections that are more consistent, easier to read and understand, and whose dependencies and other relations are made visible and available for use [1].

Similarity measures and related formalisms have been invented and reinvented many times over the years. The syntactic similarity measures we use were (re)invented by one of the authors (Alspaugh) [4], and then found to be identical with some developed by Tversky [13], who was using them in psychological research to account for how human beings perceive similarity between entities composed of parts. The same measures, as well as others more sophisticated, were already used in biology as an objective basis for classifying species more than a decade earlier [12], and by the time of Tversky's work were used in many forms in biology, paleontology, sociology, linguistics, geology, medicine, pattern recognition, criminology, and economics, among other fields [5]. Our similarity measures do not take account of the sequencing of events, but there are a number of similarity measures that do account for the sequence of elements in various ways [8]. The variety of mathematical measures of similarity is large, and the range of fields in which they are used is eye-opening; there is far too much related work to cover here, as entire papers have been

written simply summarizing the books devoted to this subject [11]. Recent examples of similarity measures include measures based on one or more ontologies [9, 10], measures that approximate the intentions of home videographers in terms of clustering of visual and temporal features of their videos [6], and measures of distance between clusters of the Web graph [7]. A good recent summary is given by Wiggerts [15], who investigates the use of similarity measures in remodularizing legacy software systems.

3 Scenario representation

We represent a scenario as a sequence of events and a set of attribute-value pairs, where an attribute may be goal, requirement, viewpoint, author, precondition, postcondition, or anything else useful for a particular scenario. Each event is further decomposed into an actor and an action, and these are represented by attribute-value pairs as well. The values for each attribute are collected into a glossary; glossary entries are reused whenever appropriate, so (for example) two events with the same actor refer to the same glossary entry. This attribute-value representation is discussed in the authors' previous work [1, 2, 4]. The representation is fundamental enough to be adapted easily to most common scenario representations. In our work, we have used it to represent scenarios that were written in prose, generally with explicit lists of events. The transformations between our attribute-value representation and prose has proven straightforward.

An example of this representation is shown in Figure 1. This scenario is identified by an identifier S_{26} and a name which are not considered in calculating similarity. The attributes of this scenario are the requirements it traces back to (3 of these), its precondition and postconditions, and the actors and actions that appear in its events. The values of this scenario's attributes would appear in five separate glossaries (Requirement, Precondition, Postcondition, Actor, and Action). The sequence of events is shown in the figure and is necessary to represent the scenario completely. Our similarity measures do not consider the sequence of events in calculating similarity, both because events are frequently transposed or reorderd in otherwise similar scenarios and in order to produce a similiarity result that is more easily comprehended. We also have not considered events themselves as attribute values in calculating similarity between scenarios, because similarity or dissimilarity of events is already accounted for by the actors and actions, albeit in a slightly different fashion.

A second example of the attribute-value representation appears in Figure 2. The values of this scenario's attributes would appear in the same five glossaries as the previous scenario's, and in fact some of the values are the same and would be referred to in both scenarios.

Attribute	Value
Requirement	R4.1
Requirement	R4.2
Requirement	R4.7
Precondition	Has (n) New
Postcondition	Has (n+1) New
Postcondition	Left Message
Actor	Caller
Actor	EMS
Action	calls the subscriber's telephone while it is busy
Action	plays the subscriber's name and announcement
Action	leaves a message

Figure 1. Scenario $\mathbf{S_{26}}$ represented as attribute-value pairs

Attribute	Value
Requirement	R4.6
Precondition	Announcement Skippable
Postcondition	Left Message
Actor	Caller
Actor	EMS
Action	calls the subscriber's telephone while it is busy
Action	begins to play the subscriber's name and announcement
Action	presses the "skip announcement" command
Action	stops playing the name and announcement
Action	leaves a message

Figure 2. A second scenario $\mathbf{S_{32}}$ represented as attribute-value pairs

An important refinement of our representation of event sequences is the use of episodes. An episode is an intentionally shared subsequence of events that may appear as part of two or more scenarios. Episodes are used to indicate dependencies between scenarios, and are somewhat analogous to subroutines. Episodes are different from unintentionally similar subsequences of events precisely in that they are intentional and represent a dependency or other relation between the scenarios in which the episode appears. In order to represent this intention, we give each episode a unique name, and scenarios refer to their episodes by name. In contrast, unintentionally similar subsequences are simply part of event sequences.

Our representation of event sequences uses a combination of sequence, alternation between two or more subsequences, iteration of potentially repeated subsequences, and episodes. The scenarios used in this study make use of an episode, but did not exhibit alternation or iteration.

4 Scenario similarity measures

A scenario *similarity measure* is a function that produces a number expressing the degree of similarity between two scenarios. In our work, a similarity measure is a function that takes two (or more) scenarios and gives a quantitative indication of their degree of similarity, ranging from 0 for completely dissimilar scenarios to 1 for scenarios identical in all respects. Within that range, a similarity measure is monotonic, never giving a lower value for scenarios that are more similar, although possibly giving the same value if the scenarios are more similar in a way that the similarity measure abstracts away from. Our similarity measures are specifically tailored for scenarios formalized as collections of attribute-value pairs.

To measure similarity, we consider each scenario as a set of attribute values. We assign the attributes embedded in episodes and events to the scenario in which they appear, so that all relevant attributes of a scenario are examined.

We define the *similarity measure* $S(S_1, S_2)$, the similarity between scenarios S_1 and S_2, as the number of common attribute values in each attribute list, divided by the sum of the sizes of each attribute list:

$$S(S_1, S_2) = \frac{2 \cdot |S_1 \cap S_2|}{(|S_1| + |S_2|)} \quad (1)$$

The factor of two normalizes the result so that identical scenarios have a similarity of 1.

As an example, we will calculate $S(S_{26}, S_{32})$ using the attribute-value pairs in Figures 1 and 2. S_{26} has 11 attribute-value pairs, while S_{32} has 10, and 5 pairs are common to both. Thus

$$S(S_{26}, S_{32}) = \frac{2 \cdot 5}{11 + 10} = .476$$

In the measure described above, each attribute value can be considered to have a "weight" of 1. A family of similarity measures arises when we allow different weights, so that each attribute is assigned a weight ranging from 0 to 1. For example, in our previous work [4], we proposed to give the "action" attribute a weight of 0 while giving all other attributes a weight of 1. Then, when the similarity measure is taken, each attribute value in the measure is multiplied by the attribute's weight. One weighting function might, for example, assign a weight of 1 to each actor, and a weight of 0 to all other attributes. With this weighting, scenarios S_{26} and S_{32} would have a similarity of .461. In this way, the weighted similarity measure emphasizes similarity in particular attributes (by assigning them high weights) and ignores difference in others (by assigning them weights of zero). This can be used for grouping similar scenarios based on particular attribute values, or for scenario searches. The weighted similarity measure can be particularly important for episode searching and matching.

Mathematically, we can write the weighted extension of S as SW, the weighted similarity measure between two scenarios, where a denotes an attribute, and $wt(a)$ denotes the weight assigned to attribute a. Note that, to avoid division by zero, we define the similarity between two scenarios to be 0 if all their attribute values have zero weights:

$$SW(S_1, S_2) = \frac{\sum_{a \in S_1 \cap S_2} 2 \cdot wt(a)}{\sum_{a \in S_1} wt(a) + \sum_{a \in S_2} wt(a)} \quad (2)$$

In the study presented here, the similarity measures used were SW with actions given weight 0, and S.

5 The plan of the study

One of the authors (Davis) produced a list of hypotheses to be examined by the study, listed below.

These hypotheses were proposed to be verified (or not) by the study:

H-1 *Distinguishing between pre- and postconditions will produce a lower similarity index than grouping all conditions together in one attribute.*

H-2 *Reconciliation of synonymous terms will more likely occur in ACTION(S), not ACTOR(S).*

H-3 *Reconciliation of synonymous terms will produce greater consistency, which will result in an overall increase in the similarity between the scenarios.*

H-4 *Intuitively similar scenarios will have a higher calculated similarity than intuitively dissimilar scenarios.*

Hypothesis 1 was added after the scenarios were initially and unintentionally cast into attribute-value form with pre- and postconditions combined in a single attribute "condition"; the goal was to learn if this accident would have made any difference in the final results.

Hypothesis 2 summarized the authors' expectations based on discussions and some previous experience in reconciling synonymous terms.

Hypothesis 4 was formulated as a minimally sufficient condition for the similarity measure to be useful in practice. This hypothesis was judged to be necessary for the measure to be of value for scenario management. As we will discuss later, evaluation of this hypothesis focused our attention on a stronger condition (which did not obtain) which is necessary for the measure to be as effective as we expected.

Similarity calculations were performed on pairs of scenarios from the Enhanced Messaging System [2]. Four sets of attribute-value pairs were extracted from each scenario, and the calculations were performed for each of the four sets. In our first paper on similarity measures, we suggested that the larger expected variability among actions might make their consideration unproductive. We have also noted the importance of appropriate reuse of values of attributes. Our representation of scenarios considers only one relationship between attribute values: they are either the same, or not the same. Therefore it is essential that values be reused if and only if they represent the same notion each time. Reconciliation of synonymous terms is necessary, else the measures will find dissimilarity where none is intended (and possibly similarity where none is intended). The four sets of attribute-value pairs reflect these views.

Set A Pre- and postconditions are considered instances of a single condition attribute

Set B No actions are considered; all other attribute-value pairs are examined.

Set C Actions are considered as well, but no reconciliation of synonymous terms is done.

Set D Actions are considered and synonymous terms are reconciled before the calculations.

Similarity calculations and evaluation of the properties and appropriate hypotheses was done for each of the sets.

The next section discusses the scenarios that were used as material for the study.

6 Material for the study

The study was conducted on a group of scenarios drawn from the Enhanced Messaging System (EMS) specification [2]. EMS is a voice mail system used by BellSouth Telecommunications for prototyping new features. Its specification consists of some 40 scenarios, some examples of which are given in Appendix A. Each scenario is characterized by one or more requirements it traces back to, zero or more preconditions, zero or more postconditions, and a list of events. Although the events are not formatted this way, each event consists of either an actor and an action, or an episode. An episode is an intentionally shared subsequence of events. One of the scenarios ($\mathbf{S_{28}}$) makes use of an episode "Make Recording" in its event list.

$\mathbf{S_1}$	Subscriber authentication
$\mathbf{S_{11}}$	Subscriber listens to a new or held message
$\mathbf{S_{12}}$	Subscriber listens to an archived message
$\mathbf{S_{23}}$	Subscriber doesn't take any action for a long time
$\mathbf{S_{26}}$	Caller calls subscriber and leaves a message
$\mathbf{S_{27}}$	Caller reviews his/her message
$\mathbf{S_{28}}$	Caller reviews and re-records his/her message
$\mathbf{S_{29}}$	Caller distinguishes his/her message as urgent
$\mathbf{S_{30}}$	Caller distinguishes his/her message as private
$\mathbf{S_{31}}$	Caller decides to speak to a receptionist
$\mathbf{S_{32}}$	Caller doesn't want to listen to the subscriber's announcement
$\mathbf{S_{33}}$	Caller calls EMS and leaves a message

Table 1. Scenarios used in the study

The scenarios are true to life in that close inspection shows that they could be expressed more clearly and consistently. The episode "Make Recording" is shared with several scenarios that were not included for this study, and it is instructive to note that although it is semantically similar to events in $\mathbf{S_{26}}$ (for example), and this similarity is intentional as far as could be determined, nevertheless $\mathbf{S_{26}}$ does not use the episode.

The scenarios are also good material for this study because the information in them is presented in such a way that it can be expressed as attribute-value pairs fairly easily, thus obviating the need to make any substantial changes that might bias the results of the study.

A list of all the EMS requirements and scenarios appears in our previous work [2]. Twelve scenarios were drawn from this collection to serve as the basis for the study. The scenarios were selected to include pairs of scenarios that appeared intuitively similar, as well as pairs that did not. The scenarios are listed in Table 1. Table 2 shows which scenarios were considered intuitively similar after detailed examination by a human analyst. The analyst considered scenarios that were not in the same group intuitively dissimilar, although this was not verified exhaustively by considering every pair separately.

$\mathbf{S_1}$, $\mathbf{S_{11}}$, $\mathbf{S_{12}}$, and $\mathbf{S_{23}}$.
$\mathbf{S_{26}}$, $\mathbf{S_{27}}$, $\mathbf{S_{29}}$, $\mathbf{S_{30}}$, $\mathbf{S_{31}}$, $\mathbf{S_{32}}$, and $\mathbf{S_{33}}$
$\mathbf{S_{28}}$ was deemed not similar to any other.

Table 2. Scenarios deemed intuitively similar by an analyst

7 Conduct of study

Organization of data and calculations on the data were performed manually by one of the authors (Davis), using spreadsheets to organize the attributes and values and to do the large amount of necessary arithmetic. The twelve scenarios were converted to attribute-value form, and glossaries for each attribute were constructed as the conversion proceeded. The attributes and values were recorded in a spreadsheet. The first two sets of data (Set B without actions and Set C with actions) were extracted directly from this spreadsheet. Set B's calculations were done by using the weighted similarity measure SW (Equation 2) with weight 0 for the actions. Set C's calculations were done using the unweighted similarity measure S (Equation 1). The use of the weight filtered out the actions for the first set of calculations.

For the third set of data, in which synonymous terms were reconciled, the analyst examined the scenarios and glossaries carefully, looking for potentially synonymous values and reconciling synonyms as appropriate. Where a value seemed inappropriate in the context in which it was used, a more appropriate value was selected and used in its place. For example, the verb "dials" was not appropriate in the action "dials the *urgent* command" because phones with dials are now uncommon, and because other similar actions used different verbs. The term "presses" was substituted for "dials" in this action and others like it. Where two terms appeared synonymous in some contexts but not in others, one term was chosen for the synonymous contexts. An example was the use of the verbs "asks" and "tells". Where no choice is possible, it is appropriate for EMS to "tell" a user something, such as "tell the subscriber to enter a subscriber's telephone number". Where a choice was given, "asks" was substituted consistently instead, as in "asks for a confirmation or rejection". Where more than one term was used for the same actor, one was chosen and used consistently. For example, "EMS" and "the system" both appeared as actors in the scenarios; "EMS" was substituted wherever "the system" had been used. Events that appeared to be incomplete expressions of an existing episode were replaced by that episode. For example, "The caller leaves a message" does not give as much information about how the message is left as the "Make Recording" episode which was already being used in similar contexts. The episode was used in place of the simple event in this case. Finally, some actions contained the word "and" and had two verbs; these actions were divided into separate events.

Similarity calculations were done for Sets B, C, and D for each of the 132 pairs of scenarios. The calculations were done in the spreadsheet and cross-checked in various ways for mistakes.

8 Study results

H-1 *Distinguishing between pre- and postconditions will produce a lower similarity than grouping all conditions together in one attribute.*

The analyst was surprised to see that the similarity index was the same in all cases; upon reflection, it was clear that the hypothesis had been carelessly framed and should have read "will produce and equal or lower similarity". Out of curiosity the analyst examined all of the EMS scenarios and found the similarity was unchanged under this distinction for for all of them. Searching farther afield she located, in an updated and unpublished scenario collection for the EMS, two pairs of scenarios for which the calculated similarity was lower if pre- and postconditions were distinguished We would expect a change in value only if one or more pre- and postconditions were the same, that is if one of them was an invariant for the scenario, and this was not the case for any of the published EMS scenarios.

H-2 . *Reconciliation of synonymous terms will more likely occur in ACTION(S), not ACTOR(S).*

One actor was reconciled but seven actions were, supporting this hypothesis.

H-3 *Reconciliation of synonymous terms will produce greater consistency, which will result in an overall increase in the similarity between the scenarios.*

This hypothesis was tested by comparing the similarities calculated on Set C to those on Set D. For example, S_{32} has three reconciled actions ("plays the subscriber's name and announcement", "stops playing the name and announcement", and "leaves a message" were each reconciled with synonymous actions from other scenarios). S_{32}'s similarity with S_{26} was 0.48 before reconciliation and 0.58 after. This pattern was repeated for other scenarios whose values had been reconciled, supporting the hypothesis. We note as an aside that reconciliation also resulted in scenarios that were more consistent, easier to read, and easier to compare by hand.

H-4 *Intuitively similar scenarios will have a higher calculated similarity than intuitively dissimilar scenarios.*

Comparison:	1	2	3
Dissimilar:	61	62	62
Mixed:	9 + 36	8 + 25	8 + 39
Similar:	26	37	23

Table 3. Scenario pairs sorted by similarity value and characterized by intuitive similarity

An initial rough validation of this was done by averaging the similarity index values for intuitively-similar scenarios and the similarity index values for intuitively-dissimilar scenarios. These average values were higher for intuitively-similar scenarios, which motivated us to examine the similarity index values in more detail. For Sets B, C, and D, the average similarity for pairs of intuitively-similar scenarios was 0.53, 0.42, and 0.43 respectively, while for intuitively-dissimilar scenarios the average was 0.19, 0.13, and 0.16 respectively. (This calculation was not performed on Set A.)

When we examined the similarity index values for individual scenarios, we noticed something interesting and unexpected. Scenario S_{28} had been deemed intuitively dissimilar to the other scenarios, yet the similarity index values indicated it was very similar to S_{27}, and markedly similar to S_{29} and S_{30}. A closer examination of S_{28} revealed that when the events of its episode were considered, it was intuitively more similar to these scenarios than had been noticed before. This was a possible example of how a similarity measure can direct an analyst's attention to where it is needed.

To obtain a more complete view of how similarity values mapped to intuitive similarity, we sorted the 132 scenario pairs by similarity value for each Set. The result was striking. In each case the lowest 61 or 62 similarity values were for intuitively-dissimilar pairs. If the pairs with S_{28} that were deemed similar after closer examination are included (S_{28} and each of S_{27}, S_{29}, and S_{30}), then the highest 26 to 61 similarity values were for intuitively-similar pairs, with a mix of intuitively-similar and -dissimilar in between. These results are summarized in Table 3, and the data for the 132 pairs and Set D is given in Table 4. It is interesting that of the pairs in the mixed stretch in the middle, all of the intuitively-dissimilar pairs involve S_{28}. If these pairs were deemed intuitively similar, the sorted pairs would break evenly into consecutive dissimilar pairs and consecutive similar pairs.

Table 4 also highlights a limitation of at least the similarity measure S (Equation 1): the calculated similarity gives no indication of where the boundary between intuitively-similar and intuitively-dissimilar lies. We discuss this further in one of the Lessons Learned in the next section.

9 Lessons learned and future work

The following conclusions were drawn from this study:

Similarity measures are useful approximations of human-determined similarity

We saw that in general, over the material examined in this study, the similarity values calculated using the similarity measure proposed in our previous work are good indications of similarity between scenarios. Even though the similarity measure only examines syntactic similarity, the results generally indicate semantic similarity as well.

Calculated similarity does not indicate the intuitively-similar-dissimilar boundary

As Table 4 makes clear, while our similarity measure gave higher values for intuitively-similar than intuitively-dissimilar scenarios, it gave no indication of where the boundary between the two groups lies. There are several distinct interpretations of this. It is possible that the similar-dissimilar distinction is an artifact of the process by which this study was set up, and that in general similarity-dissimilarity will be be a continuum, not a partitioning. It is also possible that S is not an appropriate similarity measure for scenarios, and that one that produces a larger spread of values around the boundary is more appropriate and more useful. A further study is needed to distinguish these possibilities.

The effectiveness of the similarity measure depends on glossaries and reconciliation

We saw that use of glossaries to help appropriate reuse of terms, and the reconciliation of synonymous terms, improve the effectiveness of the similarity measure.

We also saw, although the study did not address this directly, that failure to use glossaries and synonym reconciliation could cause the calculated similarity to go far astray. It was clear that without consistent reuse of terms, as encouraged and supported by glossaries, and effective synonym reconciliation, the "noise" in the calculation of similarity values could rise to damaging levels. This study confirmed our intuition that similarity measures are, in one sense, a means of capitalizing on all the many decisions about synonymity of terms that must be made during the construction of a glossary.

Better similarity values result from using more of the attributes.

The study's results indicate that the similarity values obtained from using all available attributes are generally "better" (closer to what a human analyst would find). This indicates that all the information available in a scenario is useful in determining similarity, and hints that the use of

a weighted similarity measure SW may not be more desirable than the simple unweighted similarity measure S.

Episodes complicate syntactic similarity

The empirical results were the most surprising with regard to S_{28}, "Caller reviews and re-records his/her message". This scenario, initially judged to be intuitively dissimilar to all other scenarios in the study, resulted in similarity values indicating it was similar to S_1, S_{23}, S_{26}, S_{31}, S_{32}, and S_{33}, and highly similar to S_{27}, S_{29}, and S_{30}. In examining the empirical results in retrospect, we believe this discrepancy was caused by determining intuitive similarity without expanding episodes into their component events. Thus the episode, whose five events were not shared with any other scenario in the study, "hid" five unshared attributes from the analyst.

In calculating scenario similarity, we consider each scenario to include the actors and actions of its episodes, as analysts are more interested in the similarity of two scenarios than in the similarity of their representations. We hypothesize that considering similarity with all episodes expanded into their component simple events will produce similarity values that are more consonant with intuitive similarity. We note that SMaRT provides the capability of viewing scenarios with episodes expanded, or without, and thus can give needed support in this situation.

Several areas of future work were not addressed by this study.

We showed that similarity measures usefully followed a distinction between similar and dissimilar scenarios (although without indicating it independently). A more fine-grained evaluation of similarity measures would consider how closely similarity measures mirror a range of analyst-determined similarities. For example, if an analyst ranked ten scenarios from the most similar to a given scenario down to the least similar to it, would a similarity measure produce a similar ranking? What about the more general case of ranking ten pairs of scenarios? Also, we considered our empirical results without using statistical analyses; such analysis would give results that were more broadly based, and would provide a more detailed assessment of their reliability.

Our similarity measure is chosen based on the assumption that the sequence of the scenario's events does not provide useful information about scenario similarity, and it would be instructive to compare our similarity measure with one that took account of event sequence.

We also only considered attribute weighting functions whose weights were 0 or 1, nothing in between. It is possible that graded weights might provide better results in general.

Our results showed that S did not independently indicate where the boundary between the similar and the dissimilar lies. What we do not know is whether in general an analyst perceives such a boundary; that is, whether an analyst will usually find that the scenario pairs may be cleanly divided into intuitive similar or not, or whether intuitive similarity, like calculated similarity, is a smooth gradation between "identical" and "completely dissimilar". A study with another analyst unfamiliar with similarity measures is needed.

Our results show that syntactic similarity measures can provide useful support for an analyst working with scenarios, by providing an automated indication of similarity that is a good approximation of what an analyst would find. Such similarity measures are a good choice for features of an automated scenario tool, and as a result of this study the decision was made to implement these similarity measures in SMaRT. They can provide a useful foundation from which to attack the scenario management problems that arise in the large collections of scenarios that are commonly found in practice.

A Some scenarios used in the study

S_{12}. Subscriber listens to the next message.
Requirements: R3.2.1, R3.2.2.
Precondition: $0 < s.rem$
Postcondition: $s.rem' = s.rem - 1$
1. Subscriber s dials the *next message* command.
2. EMS plays s's next message.
3. EMS changes the state of that message to 'old' if it had been 'new'.

S_{26}. Caller calls subscriber and leaves a message.
Requirements: R4.1, R4.2, R4.7.
Precondition: Has (n) New.
Postcondition: Has (n+1) New, Left Message.
1. A caller calls the subscriber's telephone while it is busy
2. EMS plays the subscriber's name and announcement
3. The caller leaves a message.

S_{28}. Caller reviews and re-records his/her message.
Requirements: R4.3.
Precondition: Left Message.
Postcondition: Left Message.
1. The caller presses the "review message" command.
2. EMS plays the message the caller just left.
3. The caller presses the "re-record message" command.
4. *Episode:* Make Recording.

S_{32}. Caller doesn't want to listen to the subscriber's announcement.

Requirements: R4.6.
Precondition: Announcement Skippable.
Postcondition: No change.

1. A caller calls the subscriber's telephone while it is busy
2. EMS begins to play the subscriber's name and announcement
3. Before the name and announcement are complete, the caller presses the "skip announcement" command.
4. EMS stops playing the name and announcement.
5. The caller leaves a message.

Episode: Make Recording

1. *Iteration with explicit exit:*
1.1. EMS tells the subscriber to begin recording.
1.2. The subscriber says what he/she wants, then presses the "stop recording" button.
1.3. EMS plays back the recording and asks for a confirmation or rejection.
1.4. *Alternation:*
1.4.1. The subscriber dials the confirmation command. *– Exit from iteration.*
1.4.2. The subscriber dials the rejection command.

References

[1] T. A. Alspaugh. *Scenario networks and formalization for scenario management*. Ph.D. Thesis, North Carolina State University, Raleigh, NC, Sept. 2002.

[2] T. A. Alspaugh and A. I. Antón. Scenario networks: A case study of the enhanced messaging system. In *Seventh International Workshop on Requirements Engineering: Foundation for Software Quality (REFSQ)*, pages 113–124, June 2001.

[3] T. A. Alspaugh and A. I. Antón. Contrasting use case, goal, and scenario analysis of the Euronet system. In *11th IEEE Joint International Conference on Requirements Engineering (RE'03)*, 2003.

[4] T. A. Alspaugh, A. I. Antón, T. Barnes, and B. W. Mott. An integrated scenario management strategy. In *Fourth IEEE International Symposium on Requirements Engineering (RE'99)*, pages 142–149, June 1999.

[5] M. R. Anderberg. *Cluster analysis for applications*. Number 19 in Probability and Mathematical Statistics. Academic Press, New York, 1973. xiii+359 pages.

[6] D. Gatica-Perez, M.-T. Sun, and A. Loui. Consumer video structuring by probabilistic merging of video segments. In *IEEE International Conference on Multimedia and Expo (ICME 2001)*, pages 709–712, 2001.

Rank	A	B	$S(A,B)$	Rank	A	B	$S(A,B)$
1.	S11	S33	0.07	67.	**S31**	**S26**	**0.21**
2.	S33	S11	0.07	68.	**S31**	**S32**	**0.21**
3.	S12	S33	0.08	69.	**S32**	**S31**	**0.21**
4.	S33	S12	0.08	70.	S1	S33	0.22
5.	S11	S26	0.08	71.	**S1**	**S11**	**0.25**
6.	S11	S32	0.08	72.	**S11**	**S1**	**0.25**
7.	S26	S11	0.08	73.	S28	S31	0.25
8.	S32	S11	0.08	74.	S31	S28	0.25
9.	S12	S26	0.09	75.	S33	S28	0.25
10.	S12	S32	0.09	76.	**S1**	**S12**	**0.26**
11.	S23	S33	0.09	77.	**S12**	**S1**	**0.26**
12.	S26	S12	0.09	78.	**S27**	**S33**	**0.27**
13.	S32	S12	0.09	79.	**S29**	**S33**	**0.27**
14.	S33	S23	0.09	80.	**S30**	**S33**	**0.27**
15.	S1	S28	0.10	81.	**S33**	**S27**	**0.27**
16.	S11	S28	0.10	82.	**S33**	**S29**	**0.27**
17.	S28	S1	0.10	83.	**S33**	**S30**	**0.27**
18.	S28	S11	0.10	84.	**S27**	**S31**	**0.29**
19.	S12	S28	0.10	85.	**S29**	**S31**	**0.29**
20.	S23	S26	0.10	86.	**S30**	**S31**	**0.29**
21.	S23	S32	0.10	87.	**S31**	**S27**	**0.29**
22.	S26	S23	0.10	88.	**S31**	**S29**	**0.29**
23.	S32	S23	0.10	89.	**S31**	**S30**	**0.29**
24.	S1	S27	0.11	90.	**S33**	**S32**	**0.30**
25.	S1	S29	0.11	91.	**S1**	**S23**	**0.30**
26.	S1	S30	0.11	92.	**S23**	**S1**	**0.30**
27.	S1	S31	0.11	93.	**S26**	**S27**	**0.32**
28.	S11	S27	0.11	94.	**S26**	**S29**	**0.32**
29.	S11	S29	0.11	95.	**S26**	**S30**	**0.32**
30.	S11	S30	0.11	96.	**S27**	**S26**	**0.32**
31.	S11	S31	0.11	97.	**S27**	**S32**	**0.32**
32.	S27	S1	0.11	98.	**S29**	**S26**	**0.32**
33.	S27	S11	0.11	99.	**S29**	**S32**	**0.32**
34.	S29	S1	0.11	100.	**S30**	**S26**	**0.32**
35.	S29	S11	0.11	101.	**S30**	**S32**	**0.32**
36.	S30	S1	0.11	102.	**S32**	**S27**	**0.32**
37.	S30	S11	0.11	103.	**S32**	**S29**	**0.32**
38.	S31	S1	0.11	104.	**S32**	**S30**	**0.32**
39.	S31	S11	0.11	105.	S28	S33	0.33
40.	S12	S27	0.11	106.	**S26**	**S28**	**0.38**
41.	S12	S29	0.11	107.	**S28**	**S26**	**0.38**
42.	S12	S30	0.11	108.	S28	S32	0.38
43.	S12	S31	0.11	109.	S32	S28	0.38
44.	S27	S12	0.11	110.	**S11**	**S23**	**0.40**
45.	S29	S12	0.11	111.	**S23**	**S11**	**0.40**
46.	S30	S12	0.11	112.	**S12**	**S23**	**0.42**
47.	S31	S12	0.11	113.	**S23**	**S12**	**0.42**
48.	S23	S28	0.12	114.	**S32**	**S33**	**0.44**
49.	S28	S23	0.12	115.	**S28**	**S29**	**0.50**
50.	S23	S27	0.13	116.	**S28**	**S30**	**0.50**
51.	S23	S29	0.13	117.	**S29**	**S28**	**0.50**
52.	S23	S30	0.13	118.	**S30**	**S28**	**0.50**
53.	S23	S31	0.13	119.	**S11**	**S12**	**0.52**
54.	S27	S23	0.13	120.	**S12**	**S11**	**0.52**
55.	S29	S23	0.13	121.	**S27**	**S29**	**0.57**
56.	S30	S23	0.13	122.	**S27**	**S30**	**0.57**
57.	S31	S23	0.13	123.	**S29**	**S27**	**0.57**
58.	S33	S1	0.15	124.	**S30**	**S27**	**0.57**
59.	S1	S26	0.17	125.	**S26**	**S32**	**0.58**
60.	S1	S32	0.17	126.	**S32**	**S26**	**0.58**
61.	S26	S1	0.17	127.	**S29**	**S30**	**0.71**
62.	S32	S1	0.17	128.	**S30**	**S29**	**0.71**
63.	**S31**	**S33**	**0.18**	129.	**S33**	**S26**	**0.74**
64.	**S33**	**S31**	**0.18**	130.	**S26**	**S33**	**0.81**
65.	S28	S12	0.20	131.	**S27**	**S28**	**0.88**
66.	**S26**	**S31**	**0.21**	132.	**S28**	**S27**	**0.88**

Table 4. Scenario pairs, sorted from least to greatest calculated similarity. Similarity was calculated including actions and after reconciliation of synonyms. The pairs in boldface are those deemed intuitively similar by the analyst.

[7] X. Huang and W. Lai. Identification of clusters in the web graph based on link topology. In *Proceedings of the Seventh International Database Engineering and Applications Symposium (IDEAS03)*, 2003.

[8] J. B. Kruskal. An overview of sequence comparison. In D. Sankoff and J. B. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*, pages 1–44. Addison-Wesley, 1983.

[9] Y. Li, Z. A. Bandar, and D. Mclean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882, July-Aug. 2003.

[10] M. A. Rodriguez and M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, Mar.-Apr. 2003.

[11] J. Scoltock. A survey of the literature of cluster analysis. *The Computer Journal*, 25(1):130–134, Feb. 1982.

[12] P. H. A. Sneath and R. R. Sokal. *Numerical Taxonomy : The principles and practice of numerical classification*. W. H. Freeman, San Francisco, 1973. xv+573 pages.

[13] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, July 1977.

[14] K. Weidenhaupt, K. Pohl, M. Jarke, and P. Haumer. Scenarios in system development: Current practice. *IEEE Software*, 15(2):34–45, Mar./Apr. 1998.

[15] T. A. Wiggerts. Using clustering algorithms in legacy systems remodularization. In *Proceedings of the Fourth Working Conference on Reverse Engineering*, pages 33–43. IEEE Computer Society, 1997.