A Qualitative Study of Similarity Measures in Event-Based Data

Katerina Vrotsou and Camilla Forsell

Linköping University, Department of Science and Technology,
Campus Norrköping, 602 74 Norrköping, Sweden
{katerina.vrotsou,camilla.forsell}@liu.se

Abstract. This paper presents an interview-based study of the definition of sequence similarity in different application areas of event-based data. The applicability of nine identified measures across these areas is investigated and discussed. The work helps highlight what are the core characteristics sought when analysing event-based data and performs a first validation of this across disciplines. The results of the study make a solid basis for follow-up evaluations of the practical applicability and usability of the similarity measures.

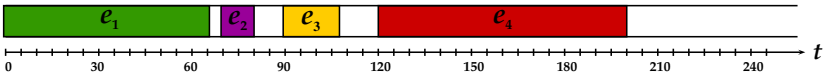
Keywords: Event-based data, event-sequences, evaluation, qualitative study, similarity measures.

1 Introduction

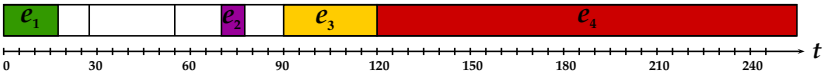
Event-based data, or event-sequence data, are encountered daily in a large range of disciplines. An event-based dataset consists of a collection of sequences of ordered events with or without a concrete notion of time [1]. This implies that the events may have an exact time-stamp associated with them or time may be relative and implicitly derived from their ordering. Examples of event-based data include medical records and procedures, time-use data, historical, biographical and career path data, internet session data, traffic incident data, process control data, and administrative process data. Even though, the data in each of these fields are of an event-based nature, they do display variations in their character. In time-use, historical, biographical and career data, for example, each recorded event has a time duration associated with it. When talking about medical event or traffic incidents the events are often regarded as instantaneous.

The process of analysing event-based data includes identifying and querying for patterns within event-sequences, comparing them and finding similarities between them. The patterns sought often have the form of shorter sequences of events, sub-sequences, which collectively exhibit a specific interesting behaviour. They may appear frequently and/or be evenly distributed, they may exhibit some sort of repetitious behaviour, or even stand out as outliers. Revealing sub-sequence similarities between data records enables their comparison, analysis and classification and can even provide a good basis for predictions.

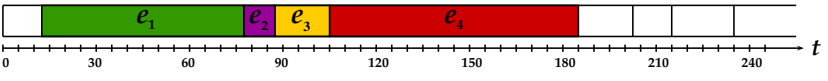
M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2011, LNCS 6771, pp. 170–179, 2011.



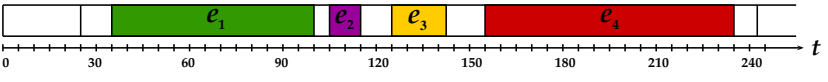
(a) Reference sequence containing the sub-sequence e_1, e_2, e_3, e_4 against which the other are compared.



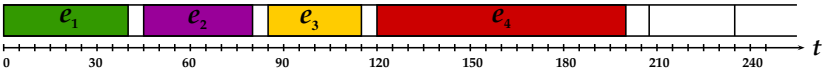
(b) The start times of the sub-sequence events are equal to the reference, their duration and intervening periods differ.



(c) The duration of each of the sub-sequence events is the same, their start times and the intervening periods differ.



(d) The sub-sequence is identical to the reference but is displaced in time.



(e) The total sub-sequence has equal duration to the reference, but the distinct events have different start times and durations.

Fig. 1. How is similarity defined? Comparison example of five event-sequences (a-e) drawn along a time-axis, t. All sequences contain the four-event sub-sequence (e_1, e_2, e_3, e_4) in different variations. Sequence (a) is the reference event-sequence against which the remaining event-sequences (b-e) are compared. The question is which sequence is more similar to the reference and why. Depending on the task and application area, different definitions of similarity are possible and hence answers may vary.

Information visualization and data-mining techniques facilitate interactive and automated analysis of event-based data by comparing and classifying individual records [1,2]. For such comparisons to be made, however, what makes sequences similar has to be determined and quantified. A set of similarity measures needs to be defined, and to do this what is characteristic, important, and/or interesting about an event-sequence has to be identified.

In previous work with time-use data, we have considered which attributes signify similarity between individuals' daily activity sequences and have identified a set of similarity measures tailored for such data [3]. In this paper we explore if and to what extent these measures have a wider applicability to event-based data. We explore four different application areas which we regard as distinct from each other and representative of several variations in the character of their data. We have performed interviews with domain experts from these areas and present and discuss the results of these interviews.

2 Motivation

Similarity as a notion can be subjective and depends largely on the characteristics of the data, the task, and the person performing the analysis. Figure 1 shows a set of event-sequences containing the same sub-sequence. The events occur in the same order but vary with respect to several measurable attributes, such as duration and start time of the sub-sequence events and duration and number of events occurring in between. The question is which event-sequences that are more similar to each other, and why. For example, when comparing work careers of individuals, the type of employment and the duration of the employment periods may be factors indicating higher similarity than the time-point during the career path when these employments started.

The motivation of this paper is to investigate how one asses similarity when presented sequences of events. What attributes define similarity in different application areas? And is there a general set of measures that can be used widely.

3 Related Work

This section will present examples of definitions of sequence similarity.

Sequence analysis is an area of high interest within computer science largely due to the extensive research in bioinformatics, which aims at identifying similarities in biological sequences based on the assumption that similarity in the structure of sequences also implies similarity in their function [4]. Techniques originally developed in this research, however, have been extended to also consider other types of event-based data. Sequence alignment [4] is such an example, in which the degree of similarity between two sequences is measured by the number of 'edit operations' (addition, deletion, substitution/move) needed to turn one sequence into the other. Each such operation carries a cost, the sum of which are referred to as edit distances, for example the Levenshtein [5] and Hamming [6] distance. Such measures define similarity based on the elements composing a sequence and the positions in which they appear. They give mostly indications of local similarity and the ordering of the elements is not directly considered. By considering sequences of elements, segments of a sequence, instead of single elements, in the similarity computation the ordering of the elements can be given importance, as in [7].

Similarity between event-sequences has been researched by Moen [8,9]. She too defines similarity based on edit operations but extends these to incorporate the occurrence time of the events [8]. This is done by assigning costs depending on the number of occurrences of an event type as well as how far an event is moved (time difference). Therefore emphasis is given to the importance of both event type and timing when comparing sequences. Furthermore, the concept of similarity based on context of events is introduced [8,9]. The context of an event is defined as the set of all event types that occur within a certain time frame before it. With this extension focus is put, not only on the event types and order of a sequence, but also on the relationships between event types and their importance in the structure of a sequence.

Matches and mismatches of events in compared sequences are considered in the sequence similarity measure presented in [10] which brings attention to the events interrupting a sequence of interest, meaning events that occur in between the matched events. This measure defines similarity in terms of event types, order and relative duration and is extended to also consider number of mismatched events, meaning the number of interrupting events [11]. Another explored measure of similarity between sequences is the frequency of occurrence of segments of consecutive elements, sub-sequences [12]. Such measures are based on the assumption that similar sequences share common sub-sequences. Similarity can thus be defined based on the number of common events and the ordering of these events, as in [13] where a similarity measure for categorical data is presented.

As evident from above, similarity between sequences can be based on many characterizing attributes which depend on the type of data as well as the objective of the task. Often, presented similarity measures are followed-up by evaluation studies performed using a certain type of data. These measures, however, are not as often evaluated across diverging application areas.

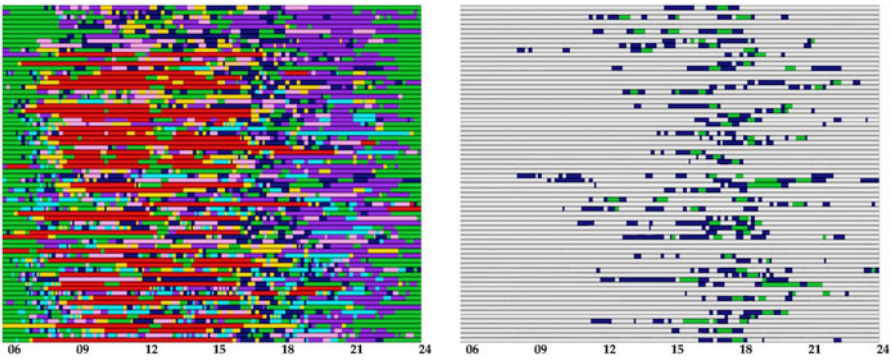
4 The Study

Nine similarity measures have been previously identified for comparing activity diary data [3]. An activity diary is composed of a sequence of activities performed during the course of the day. Within each diary, there are sequences of activities performed in order to achieve certain tasks (activity projects). For example, 'grocery shopping', 'preparing food', 'eating', and 'doing the dishes' can be regarded as parts of the activity project *'have dinner'*. Looking at how such sub-sequences are incorporated across populations (figure 2(b)) reveals similarity between individuals and facilitates the extraction of similarly behaving groups.

Since activity diaries belong to the larger category of event-based data, it is interesting to investigate whether the identified measures of similarity are widely applicable to this larger category. For this reason we have investigated similarity in the following four event-data types representing different application areas:

- *Time-use data*; activity diaries, biographies, careers.
- *Medical health records*; patient journal data.
- *Process control data*; air traffic control, electricity flow, paper mill industry.
- *Administrative process data*; patient administration, medical process data.

These categories are chosen because each of them represents a variation of event-based data. In time-use data events have a start time and a duration and each event starts immediately after the previous. Medical health records have a time-stamp associated to each event but these do not necessarily have a duration. Also, the periods intervening events (interruptions) are of importance, especially the duration of these. Web session data and traffic incident data are also comparable to this type of data. In process control data, instead, a continuous flow is usually observed and the events have the form of abrupt changes in this flow. Administrative process and industrial process data are similar in that they are defined as



(a) Event-sequence dataset (b) Sub-sequence highlighted

Fig. 2. Example of a sub-sequence highlighted within a dataset of event-sequence records. Similarities in how the sub-sequence is incorporated and distributed in the data become apparent. Figure (a) shows a population of women aged 25 and older and (b) the distribution of the activity project *'have dinner'* within this population. This activity project includes the activities: 'grocery shopping', 'preparing food', 'eating', and 'doing the dishes'.

processes of events that should occur and/or followed. Event-sequences reflecting how these pre-defined sequences are executed in reality can be collected in order to improve performance.

The measures that have been identified as representative for making such comparisons in activity diaries, and that are investigated here for their general applicability are listed (enumerated) below. They are concerned with the general character of each record (1-2), with the identified sub-sequence characteristics (3-6), and with the events interrupting an identified sub-sequence (7-9). In order to best explain these measures we use an example. Consider the event-sequence of figure 3. The measures will be defined by looking at how the sub-sequence 'ABC' is identified within the total event-sequence.

1. *Fragmentation* refers to the number of events composing an event-sequence; the length of a sequence. In the example of figure 3 fragmentation is 11.
2. *Variation* refers to the number of unique events composing the event-sequence; the number of distinct event types present. In the example of figure 3 the variation is 5 (A, B, C, D, E).
3. *Number of occurrences* is the number of identified occurrences of an event or a sub-sequence, meaning the number of matches found within the total event-sequence. There are 4 instances of 'ABC' in the example of figure 3.
4. *Start time* refers to the initiation time of an identified sub-sequence, meaning the start time of the first event of the sub-sequence. The start times for each of the four identified instances of 'ABC' are $< 0, 0, 18, 23 >$.
5. *Sequence length* refers to the total duration of an identified sub-sequence, from the start time of the first matched event until the end time of the last

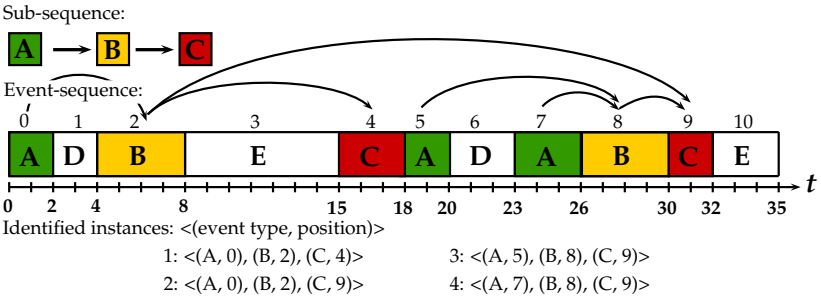


Fig. 3. Example of sub-sequence A, B, C identified within an event-sequence with 10 elements along a time-axis, t, going from $0 - 35$. Four instances are identified.

one including any possible interrupting events. The sequence length for each of the four identified instances of 'ABC in figure 3 are $< 18, 32, 14, 9 >$.

6. *Sequence event length* refers to the sum of the durations of the distinct events composing the identified sub-sequence. The sequence event length for each of the four identified instances of 'ABC' in figure 3 are $< 9, 8, 8, 9 >$.
7. *Number of interruptions* refers to the number of events interrupting an identified sub-sequence, the number of intervening events. The number of interruptions for each of the four instances of 'ABC' in figure 3 are $< 2, 7, 2, 0 >$.
8. *Length of interruptions* refers to the sum of the durations of the events interrupting an identified sub-sequence. The length of the interruptions for each of the four instances of 'ABC' in the example of figure 3 are $< 9, 24, 6, 0 >$.
9. *Type of interruptions* refers to the type of the events interrupting an identified sub-sequence, the context of the interrupting activities. In the example of figure 3 one event of type D and one of type E are interrupting the first identified instance of 'ABC'.

These nine measures cover several aspects concerning the measurement of similarity between event-sequences and we believe them to be widely applicable. They were, however, developed for a certain type of data and area. The objective with the present study has, therefore, been to explore their potential use and applicability for data analysis in other areas as well.

4.1 Method

This section describes the method applied and the results of the study .

Participants. Eight participants took part in the study, five female and three male aged between thirty and fifty-nine years. The participants were experts in the areas of time-use data (two), process-control data (two), administrative process data (two) and medical data in the form of electronic health records (two). All of them were potential users of similarity measures in their current work, however, none of them were familiar with the idea or the measures prior to the study. They received no compensation for taking part in the study.

Material and Procedure. Each participant was interviewed individually. Each session was performed as a one hour (approximately) discussion around a semi structured interview guide. This guide included twelve predefined questions and a number of potential questions used to prompt discussion. The goal was to explore the following question themes for each of the four application areas:

1. The *current process* of collection and analysis of event-based data.
2. The *attributes of interest* in the search for sequence similarity.
3. The *definition of similarity* between sequences.
4. The *applicability* of the nine proposed measures.
5. The *benefit* of the measures in terms of improving data analysis.
6. Any *additional measures* of similarity useful for data analysis.

Some days prior to the interview session each participant was provided with a written document for information and preparation. The document included a description of the background and purpose of the area and the study including descriptions of the measures and illustrations similar to figures 1 and 3. The interview session began with a joint review and discussion of this document and the interviewer provided further explanations when needed. This was done in order to facilitate the understanding of how the measures were defined and to spark the participants' reasoning and imagination on how how they could be applicable in that persons actual area of expertise.

4.2 Results

A summary of the interviews' results is presented in this section.

Current process: Two main types of analysis of processes were identified. In time-use data and in medical health records a sequence of events consists of records from a person. These are events of an arbitrary nature, number and order. The major analysis task is to compare data records of different persons, or previous data records of the same person. In process control and administrative process data a sequence is defined by a chain of predefined events. Such a sequence can be regarded as a reference and any actual outcome of a process can be compared with its ideal result.

Attributes of interest: Common for all areas was that the importance of the attributes depends on the task at hand. Occurrence and duration was recognized as important for all application areas. For time-use data and medical health records the start time of an event-sequence was critical, for example, an open wound must be sawn within a certain time frame from when it appeared. For process control and administrative processes time is also a critical attribute, not the actual start time of an event, however, but rather the duration of a sequence after it has started. See also table 1.

Definition of similarity: Here the main finding concerns the order of events as a relevant factor for defining similarity. This factor certainly has an impact on what attributes, and thus what measures, that are of interest. As mentioned above, for data in administrative processes and in process control there already exists a predefined order of events so this factor is inherent for any sequence.

Table 1. Applicability of similarity measures to the four application areas. Applicability is indicated by 'x' and inapplicability by '-'.

	time-use data	medical records	process control	administrative process
1. Fragmentation	x	x	-	-
2. Variation	x	x	-	-
3. Number of occurrences	x	x	x	x
4. Start time	x	x	-	-
5. Sequence length	x	x	x	x
6. Sequence event length	x	x/-	x	x
7. Number of interruptions	x	x/-	x	x
8. Length of interruptions	x	x	x	x
9. Type of interruptions	x	x	x	x

Hence, similarity is defined as consistency with a fixed reference sequence. For time-use data and medical health records, on the other hand, order is of great importance since different order can indicate different behaviour. The definition of similarity depends also on the definition and composition of the event-sequence. In most areas there are several ways to define the same sequence with respect to the type of considered events, their description detail, and the considered time unit. For example, in medical health records a sequence can be defined to include doctor visits, lab tests, diagnoses, and treatments, alternatively it can also include symptoms, treatments and recovery periods. In the first case the events are considered instantaneous while in the second case events have a duration which is of importance. The specification hence of the attributes that indicate similarity depends on the sequence composition.

Applicability: The applicability of the nine measures is illustrated in table 1. For time-use data the applicability of all measures was confirmed. For medical health records the applicability of two measures was variable depending on the definition of what constitutes a sequence. The first is *'Sequence event length'*. If the analysed sequence considers events instantaneous (as defined previously) then it is intrinsic that the events have no length, so the measure is inapplicable. The second is *'Number of interruptions'* which depends on the definition of the sequence and on who is in focus in the analysis. When looking at medical events as durable entities then the number, duration and type of intervening events are important while otherwise the duration that a patient has to wait between events is of interest. Also, the applicability of this measure may vary depending on who is in focus. For the patient, for example, the waiting time between events (duration of interruptions) is important, while for the doctor it may be the type of these interrupting events. In medical health records also *'Start time'* is extremely important since it dictates choice and effect of various treatments. In administrative processes, on the other hand, time is a relative attribute. Tasks have to be handled within a certain time frame after their initiation so the duration of an event or sub-sequence is of importance while *'Start time'* is irrelevant. The process control area shows the same applicability

pattern as administrative processes. In general, when a sequence of events is given beforehand neither *'Fragmentation'* nor *'Variation'* are relevant. Furthermore, all aspects of interruptions are of importance in any such process.

Benefit: Much attention was focused towards optimization in all application areas. In process control and administrative processes the measures allow various comparisons of an actual outcome with a reference outcome. Hence one can detect what and where in a sequence deviations and anomalies from that idealized result are present. Here the measures can be used as tools for reflection to optimize behaviour, procedures and actions to reach a desired outcome. For time-use data and medical health records the main benefit was identified as an enhanced ability to find, identify, cluster and compare important occurrences, patterns and trends of events or sequence leading to improved analysis. Also, all participants expressed that the measures have a wide use when it comes to mining and fusion of statistics e.g., frequencies, durations and other types of various attributes of similarity.

Additional measures: In time-use data and medical health records it is interesting to compare an identified sequence of events as a whole, whereas in process data it is more important to study the characteristics of the distinct events composing the identified sequence, such as when an event starts and how long it lasts. As a result, a new measure was identified as *'Event start'* that will measure and compare the starting point of each distinct event in a sequence. This start can either be an exact time point, or the starting point in relation to the beginning of the sequence; the second day after the initiation of a task, for example. The measure is applicable to all considered areas.

5 Discussion and Conclusions

This paper is part of a larger effort to develop similarity measures to support analysis of event-based data. We have validated nine measures for time-use data and shown that they are general enough to also be applicable to other areas. The results, however, highlighted the need for careful definition of what constitutes an event-sequence since there are several possible options for each area and the applicability of the measures may vary accordingly. A summarizing conclusion is that event-based data often represent two categories of sequences.

1. *Either things happen*; events or incidents occur, activities are performed, sequences of choices are documented, and then one looks at how, when and why they happen. In general, data that represent events that are performed 'freely' fall under this category, for example activity diaries, electronic health records, incident data, web sessions and transaction data over time.
2. *Or things should happen*; sequences of events representing a process are predefined and should be followed to perform a certain task, and then one examines how close the actual outcome is to this optimally defined sequence. Data that represent the turn out of a predefined sequential process, may that be administrative or industrial, fall under this category.

For both categories, order, frequency, time and interruptions are all relevant attributes in their different forms since they are intrinsic characteristics of a sequence. Measures such as fragmentation and variation are applicable to the first category of data but not to the second category where event-sequences are predefined. Also start time is irrelevant in this case as a measure since it is the total duration of a process that is important, not when the process was initiated.

Further, we have identified an additional attribute to consider when defining similarity, namely the starting point of each event in a sequence. This resulted in the new measure *'Event start'* which is applicable to all areas.

To conclude, the study has confirmed the theoretical applicability of the measures with respect to what an analyst is interested in when comparing event-sequences. These results will provide solid basis for forthcoming development of the measures and for assessment of their use and usability in practice with quantitative evaluations.

References

1. Han, J., Kamber, M.: Data mining. Concepts and techniques. Morgan Kaufmann, San Francisco (2006)
2. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
3. Vrotsou, K.: Everyday mining: Exploring sequences in event-based data. PhD thesis, Linköping University (2010)
4. Durbin, R., Eddy, S.R., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press, Cambridge (1998)
5. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707–710 (1966)
6. Hamming, R.W.: Error detecting and error correcting codes. Bell System Technical Journal 26(2), 147–160 (1950)
7. Ergun, F., Muthukrishnan, S., Sahinalp, S.C.: Comparing Sequences with Segment Rearrangements. In: Proceedings of Foundations of Software Technology and Theoretical Computer Science, pp. 183–194. Springer, Berlin (2003)
8. Moen, P.: Attribute, Event Sequence, and Event Type Similarity Notions for Data Mining. PhD thesis, Dept. of Computer Science, University of Helsinki (2000)
9. Mannila, H., Moen, P.: Similarity between Event Types in Sequences. In: DaWaK 1999: Proc. of the First International Conference on Data Warehousing and Knowledge Discovery, Florence, Italy, pp. 271–280. Springer, Heidelberg (1999)
10. Wongsuphasawat, K., Shneiderman, B.: Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In: IEEE Symposium on Visual Analytics Science and Technology, pp. 27–34 (2009)
11. Wongsuphasawat, K., Plaisant, C., Shneiderman, B.: Querying Timestamped Event Sequences by Exact Search or Similarity-based Search: Design and Empirical Evaluation (2010)
12. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence comparison.. Proc. of the National Academy of Sciences of the USA 85(8), 2444–2448 (1988)
13. Gómez-Alonso, C., Valls, A.: A Similarity Measure for Sequences of Categorical Data Based on the Ordering of Common Elements. In: Modeling Decisions for Artificial Intelligence, vol. 1, pp. 134–145. Springer, Heidelberg (2008)