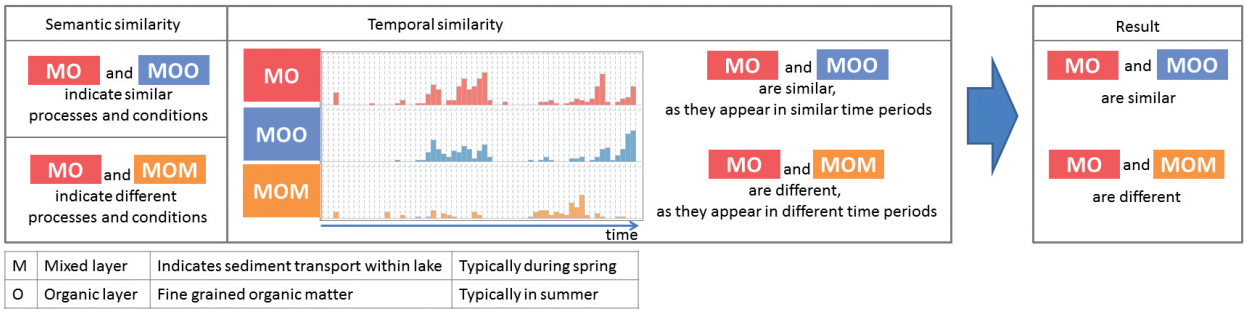
Understanding a sequence of sequences: Visual exploration of categorical states in lake sediment cores

Andrea Unger, Nadine Dräger, Mike Sips, and Dirk J. Lehmann



M	Mixed layer	Indicates sediment transport within lake	Typically during spring
O	Organic layer	Fine grained organic matter	Typically in summer

Fig. 1. The main analytical task is to determine which categorical sequences in the time series are similar. To this end, geoscientists assess semantic and temporal similarity. The "semantics" of a categorical sequence denote the geoscientific meaning, which is interpreted based on domain knowledge. The temporal similarity is assessed from the time periods in which the sequences appear.

Abstract—This design study focuses on the analysis of a time sequence of categorical sequences. Such data is relevant for the geoscientific research field of landscape and climate development. It results from microscopic analysis of lake sediment cores. The goal is to gain hypotheses about landscape evolution and climate conditions in the past. To this end, geoscientists identify which categorical sequences are similar in the sense that they indicate similar conditions. Categorical sequences are similar if they have similar meaning (semantic similarity) and appear in similar time periods (temporal similarity). For data sets with many different categorical sequences, the task to identify similar sequences becomes a challenge. Our contribution is a tailored visual analysis concept that effectively supports the analytical process. Our visual interface comprises coupled visualizations of semantics and temporal context for the exploration and assessment of the similarity of categorical sequences. Integrated automatic methods reduce the analytical effort substantially. They (1) extract unique sequences in the data and (2) rank sequences by a similarity measure during the search for similar sequences. We evaluated our concept by demonstrations of our prototype to a larger audience and hands-on analysis sessions for two different lakes. According to geoscientists, our approach fills an important methodological gap in the application domain.

Index Terms—Visualization in Earth Science, Time Series Data, Categorical Data, Design Study.

1 Introduction

Understanding landscape development, the function of ecosystems, and the responses to climate change is of fundamental interest to geoscientists. The investigation of developments in the past is valuable to assess today's and future's climate dynamics and their effect on landscape evolution. Geoarchives are valuable sources in this regard, as they chronologically capture responses of the ecosystem to climate and environmental conditions. The responses result from an interplay of multiple local and global driving factors including not only climate, but also local geology and geochemistry, vegetation, animals and human impact. Deducing the conditions from the responses requires domain knowledge. To identify and disentangle the factors, multiple types of geoarchives from numerous locations are investigated. Furthermore, hypotheses gathered from one study are tested in others. Each geoarchive is therefore one piece that contributes to the overall picture.

In this work, we focus on the analysis of one important type of geoarchives: sediment cores which are recovered from the ground of lakes. Microscopic analysis leads to so-called "microfacies data", a time sequence of categorical sequences. One categorical sequence comprises the sediment layers in one year. The geoscientific goal is to gain hypotheses from this data about the climate and environmental conditions that appeared in the past and about their temporal extent. Geoscientists approach this goal by identifying groups of categorical sequences that indicate the same conditions. To this end, the geoscientist assesses the semantics and the temporal similarity of sequences. Sequences are semantically similar if they have similar geoscientific meaning. They are temporally similar if they appear in similar time periods. An example is shown in Fig. 1. Assessing the similarity of categorical sequences becomes a challenge when the data set comprises hundreds of unique sequences and shows high temporal volatility.

Our main contribution is a tailored visual analysis concept that supports the investigation of microfacies data. Visual methods support the simultaneous assessment of the semantics and the temporal context of categorical sequences. Computational methods substantially facilitate the identification of similar categorical sequences in the data. They automatically identify unique sequences and provide a ranking during the search for similar categorical sequences. Interactive means support the generation and adjustment of groups effectively. Our concise visual interface allows the user keeping track of the analytical progress.

This design study was conducted in close collaboration between researchers from visualization and geoscience. An expert for microfacies analysis co-authored the paper. She accompanied the scientific process, which comprised gaining an initial understanding of the geoscientific question (Sec. 2) and the analytical procedure with the main tasks (Sec. 3), the conceptual development (Sec. 5) and its realization with specific

- *Andrea Unger, Nadine Dräger and Mike Sips are with GFZ German Research Centre for Geosciences. E-mail: {andrea.unger, nadine.draeger, mike.sips}@gfz-potsdam.de*
- *Dirk J. Lehmann is with University of Magdeburg. Email: dirk@isg.cs.uni-magdeburg.de*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

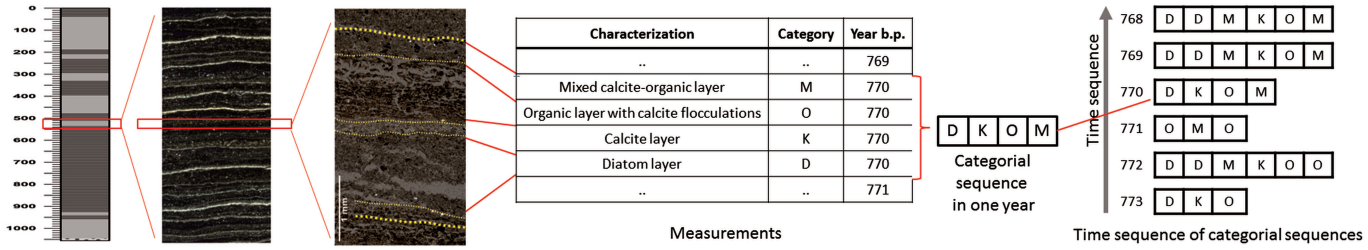


Fig. 2. Microscopic analysis of a sediment core. The sediment core (on the left) comprises layered sediments over thousands of years. Magnification under the microscope provides the necessary detail to record qualitative measurements. The yellow lines indicate the boundary between different layers, whose qualitative characterizations are depicted in the table. A categorical sequence subsumes the sediment layers that correspond to one year. The result is a time sequence of categorical sequences. Years b.p. denote years before present (the reference for present is 1950).

methods (Sec. 6 and 7). In addition, five microfacies experts confirmed that the analytical goals are relevant for their work and that they apply the same analytical procedure. The resulting prototype was evaluated (Sec. 8) in use cases by two microfacies experts, one of them being the co-author. Further feedback was gathered from a demonstration of the prototype to eight researchers from a leading group in climate and landscape development. Our prototype was ascribed as the first systematic method for the analysis of microfacies data, which fills an important methodological gap in the application domain.

2 Geoscientific Background

2.1 Microfacies analysis of sediment cores

Lakes are highly sensitive to changes of climate and environmental conditions. Annually laminated lake sediments reveal unprecedented details, as the conditions are reflected in seasonal layers [7, 9, 28]. Microscopy of the seasonal layers (microfacies analysis) is one approach that is employed to extract information about past climate and environmental conditions [6]. During microscopic analysis, geoscientists describe the characteristics of the individual layers [5]. Usually, multiple layers deposit during a year. They reflect the annual cycle of the lake's ecosystem with its variations, e.g., in water circulation, plant growth, or temperature. Based on these annual cycles, the geoscientist subsumes seasonal layers into intervals that correspond to one year. Note that more than one layer can be deposited during a season, or that a specific season may not be represented by a layer, as no sediment was deposited in the season (i.e., in winter). The dating of years is derived by counting the annual cycles backwards from present.

2.2 Microfacies data

The result of microscopic analyses is a time sequence of categorical sequences (see Fig. 2). The overall time sequence comprises thousands of years. Each year is described by a categorical time sequence. A categorical sequence is formed by the sediment layers in one year. Each sediment layer represents a single categorical state. The states within a categorical sequence do not have an absolute time stamp. Solely their relative temporal order is known.

2.3 Analytical goals: conditions, responses, and temporal extent

Geoscientists aim at deducing hypotheses about developments of landscape and climate from microfacies data. They investigate two main questions:

- Which climate and environmental conditions appeared in the past and what was the response of the landscape?
- When did the climate and environmental conditions appear? What was their temporal extent?

Fig. 3 illustrates the goal. On top, the available data is shown: the time sequence of categorical sequences. Below, we see the result that geoscientists are looking for: Climate and environmental conditions (A, B, C, D in the example) and their temporal extent (time periods T_1 to T_6) are deduced. Note that the temporal extents of conditions strongly vary, even if the same condition reoccurs in different time periods.

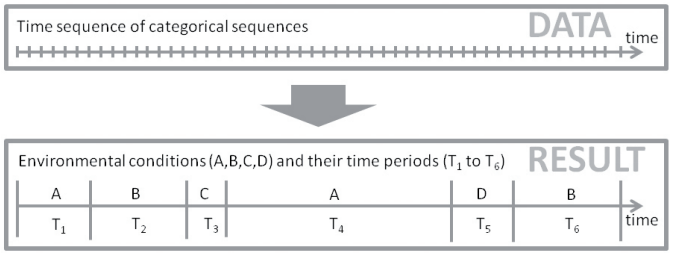


Fig. 3. Analytical goal: The time sequence of categorical sequences (on top) is analyzed to identify the climate and environmental conditions in the past (A, B, C, D) and their temporal extent (time periods T_1 to T_6).

3 Analytical Procedure and Challenges

3.1 Analytical Procedure: Find groups of related sequences

Geoscientists apply the following approach to gain hypotheses about climate and environmental conditions in the past and their temporal extent: They derive groups of categorical sequences that represent responses to the same conditions. The categorical sequences within a group provide indications about the climate and environmental conditions and the responses of the landscape. The temporal extent of conditions is given from the occurrences of the group over time. To find meaningful groups, it is necessary to determine which categorical sequences are responses to similar conditions and which are responses to different conditions. Therefore, the analyst applies three steps which we describe in the following.

3.1.1 Identify unique categorical sequences

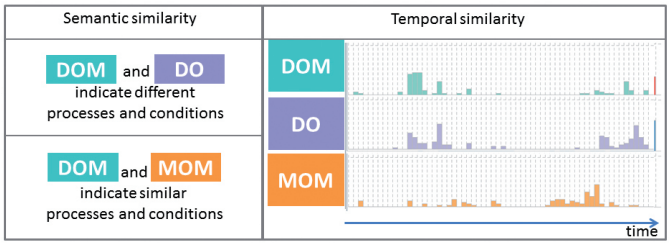
Commonly, multiple time points exhibit the same categorical sequences. Same categorical sequences point to the same climate and environmental conditions. To gain an overview which categorical sequences appear in the data, the set of unique sequences is identified. Further, the temporal occurrences of each unique sequence are inspected to understand which time points indicate the same climate and environmental conditions.

3.1.2 Group similar unique sequences

The same climate and environmental conditions can cause different responses. This results in different categorical sequences that represent the same conditions. In consequence, the unique sequences are explored to identify which sequences indicate the same climate and environmental conditions. Unique sequences are considered as indications for the same conditions if they have similar semantics and similar temporal context. We explain both in the following.

Semantic similarity. The term "semantics" refers to the meaning of a categorical sequence. Understanding the meaning of a categorical sequence requires domain knowledge. In our geoscientific application, a categorical sequence is an indicator for certain climate and environmental conditions and the responses of the landscape. Categorial sequences are semantically similar if they indicate the same conditions and responses. In Fig. 1 and Fig. 4, we present examples for semantic similarities and differences. The examples show that it is not sufficient to compare the sequences of states. All categorical sequences show strong similarities. But the analyst considers some of them as similar and some of them as different, based on domain knowledge. Also, it is common that categorical sequences appear in the data whose meaning is not yet fully understood.

Temporal similarity. Climate and environmental conditions appear in delimitable time periods. In consequence, categorical sequences that indicate the same conditions also occur in delimitable time periods. Categorical sequences are temporally similar if they appear in the same time periods. Hence, the analyst investigates the temporal occurrences of categorical sequences. Fig. 1 and Fig. 4 provide examples of temporal similarities and differences among categorical sequences.



M	Mixed layer	Indicates sediment transport within lake	Typically in spring
O	Organic layer	Fine grained organic matter	Typically in summer
D	Diatom layer	Indicates algae bloom in the lake	Typically in spring

Fig. 4. Assessment of the semantic and temporal similarity for two pairs of categorical sequences. The similarity assessment leads to ambiguous findings: The sequences *DOM* and *DO* have different semantic meaning. The *M* layer is an indicator for high circulation, while the absence of that layer in *DO* indicates low circulation. But they appear in similar time periods. The second pair of sequences, *DOM* and *MOM*, have similar semantics, but they are not temporally similar. The geoscientist concludes that neither *DO* nor *MOM* are similar to *DOM*.

The examples in Fig. 4 show that it is important to concurrently assess semantic and temporal similarity of categorical sequences, as both can lead to different results.

3.1.3 Inspect groups of related sequences

During analysis, the data set is transformed from a time sequence of categorical sequences into groups of unique sequences. Geoscientists inspect these groups and their temporal occurrences to derive hypotheses about landscape and climate conditions in the past.

3.2 Analytical challenges

At the heart of the analytical procedure, the meaning of differences among unique sequences is assessed as a prerequisite to find meaningful groups of related sequences. This is carried out by comparisons of semantics and temporal context of unique sequences. Both types of comparisons rely on human assessment. Interpreting the meaning of categorical sequences involves domain knowledge. Determining the temporal similarity of unique sequences also uses human assessment, as criteria to determine temporal similarity are not known a-priori. Conducting these comparisons becomes a challenge if the data set comprises many unique sequences and is strongly volatile over time. The exemplary data set which we investigate in Sec. 8.1.1 comprises around 600 unique sequences over more than 6,000 years.

The main goal of our work is to provide methods that systematically support this analytical procedure. We specifically focus on the challenges that come with many unique sequences and high temporal volatility.

3.3 Available Methods for Microfacies Data

So far, our domain experts have interpreted sediment core data individually and manually, based on a visual inspection of the data. During microscopy process, they develop hypotheses about temporal developments and typical characteristics of states and sequences. These hypotheses are examined with statistical tests and graphical data plots. Available methods to identify temporal evolution do not consider the volatility of sequences of states as a criterion. Inspections of categorical sequences focus on short time ranges, which is plausible as investigations are time consuming. The temporal course of categorical sequences in data sets with many different sequences is rarely investigated due to a lack of methods that support this task. Our collaboration partners are aware of this aspect and asked for novel analytical approaches that support an investigation of categorical sequences across thousands of time points and speed up the process. Hence, our approach is tailored to enable a systematic analysis of the data on a high-quality level via an interactive visual analytics framework and in acceptable time.

4 Related Work

Due to the interdisciplinary nature of our application issue, related work brushes various areas. We discuss relevant mining approaches as well as visual data analysis of temporal data and sequences.

4.1 Mining Techniques

The basic conditions in order to consider state-of-the-art mining techniques are – in terms of our scenarios – two-fold: (i) the similarity of sequences and (ii) the relation of sequences in time. Finding sequences of interest in time is related to time-dependent distribution properties of sequences. Regarding this, mining algorithms consider data as single sequences of categorical states. The aim is to detect potentially interesting motifs, which are subsets of the categorical states. Similarity trees [32] could support to find relevant patterns, but lack to involve the expert's domain knowledge, while correlation based approaches [45] lack to integrate the time aspect appropriately. Pixel-based bar chart techniques [22] seem to be a good starting point to aggregate data. Our approach goes in a similar direction. Nevertheless, the detection of motifs is based on pre-defined assumptions. In that regard, a number of mining algorithms detect specific patterns, e.g., such as periodic patterns [11,44] or surprising patterns [1] (= outliers). The challenge with automated algorithms is that our users do not know in advance what constitutes an interesting pattern.

Closely related to our work, a popular approach is to group similar elements to clusters [1,15,34,35,46]. For doing so, relevant geometric, geodesic, or stochastic distance measures need to be defined in the first place, as criteria to group elements together. Since categorical data cannot be considered to be embedded within an Euclidean space, this task is a challenge. In high-dimensional spaces, the contrast for clusters vanishes [4,18], rising questions to cluster reliability. Further, similarity among categorical sequences in our application is not fully explained by the similarity of categorical states (which can be captured by formal similarity measures), but emerges from interpretation of sequences based on domain knowledge.

4.2 Visual Analysis of Sequence Data

Time sequence analysis is a major interest of the visualization community. Numerous methods have been presented to handle temporal data (overviews in [2,37]). Often, numerical data is in the focus, whose analysis demands other operations and visualization methods than the categorical sequences in our application domain. Still, several works on numerical time sequence data inspired our research. Multi-resolution techniques [16] employ the concept of arranging time in more than one spatial dimension, which is adapted in our visual design. The TimeClassifier [41] supports the identification of relevant patterns in long time series with semi-automatic approaches and user-driven grouping. Our strategy is similar, but our time sequence of categorical sequences

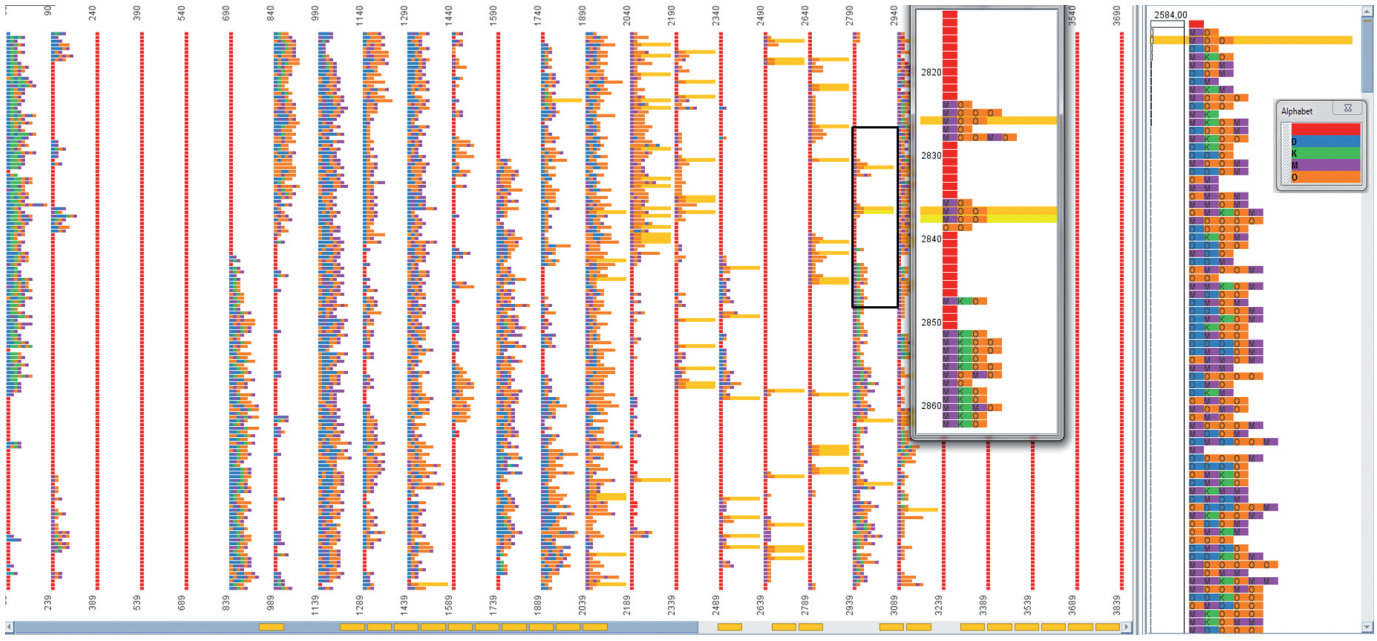


Fig. 5. Visualization for Step 1: Identify unique sequences and their temporal occurrences. The semantic view (right) shows the unique sequences sorted by frequency from top to bottom (most frequent on top). Each row depicts a unique sequence. Each unique categorical sequence is shown as a series of colored blocks from left to right. A block's color indicates the state. Frequency is explicitly visualized in small bar charts left to the sequences. The temporal view (left) shows a highly compact representation of the complete time sequence of categorical sequences. The time sequence is split into multiple columns, which are aligned side by side. In a single column, time is mapped from top to bottom. Each row shows the categorical sequence at one time point by diminished visual primitives compared to the semantics view. Small time frames can be shown in detail on the user's demand (the time subsequence within the black frame is magnified in the detail view on top). Both views are coordinated with a brushing and linking mechanism: The semantic view highlights a selected unique sequence of interest and the temporal view the corresponding time points.

requires a different assessment of similarities. Further, interaction techniques for time sequences [20, 40] are relevant for our work.

Over the last years, visualization research addressed categorical data. Categorical sequences consist of events or states. Their elements may be associated with time stamps, turning the categorical sequence into a time series. An early contribution was Life Lines [33], which solely focused on a single record. Finding similar temporal categorical records was proposed in [43]. In LifeFlow [42] and EventFlow [30], events within sequences are visually grouped to reveal similarities and differences in sequences and to enable sequence simplification strategies. More general, strategies to handle high volume and variance in categorical time sequences are presented in [10]. All these methods dominantly work on sets of categorical time sequences. Note that our times series have a different interpretation: we have a time stamp per categorical sequence, but not for the elements within. Hence, these methods are not applicable to our problem, as the temporal relation among categorical sequences is a crucial part of microfacies analysis.

In addition, there are techniques dealing with categorical data in different application domains. They inspired and influenced our work. To mention a few: In the domain of human geography, Vrotsou et al. [39] tracks similarly behaving records based on user-input. Albers et al. [3] address genomic alignment data, while Coco [27] is designed to compare two disjoint sets of records. Again, the focus lies on analyzing a set of categorical sequences rather than on understanding temporal relations among them. Similar subsequences within one long time sequence are studied in [38]. An alternative to showing categorical sequences explicitly could be low-dimensional embeddings [19, 21, 23, 24], but a semantic equivalent map from our categorical data to numerical data does not exist.

To conclude, traditional approaches do not allow an exhaustive analysis of sediment data. Regarding this, our domain experts have a strong expertise in geoscience, while we are experts in (visual) data analysis. Thus, the main task is to combine the expertise of the different disciplines and facilitate an interactive visual access to the data for our experts. In fact, there is a need for a visual analytics approach to analyze sediment data, which motivates to fill this gap in the literature.

5 Visual Analysis Concept

In this section, we introduce our visual analysis concept to support the analytical procedure introduced in Sec. 3. We show the interplay of visual, interactive, and computational methods for each step. Our visual design is explained in more detail in Sec. 7.

5.1 Identify unique categorical sequences

The initial step of the analytical procedure is to identify unique categorical sequences throughout the time sequence. We apply computational methods to extract the unique sequences in the data set and link them to time points. Our visualization (Fig. 5) comprises an overview on the semantics and an overview on temporal developments. The semantic overview displays all unique sequences and their frequency. The time overview shows the time sequence of categorical sequences in a highly compact fashion and provides details on demand. Both views are coordinated with a brushing and linking concept that supports the inspection of single unique sequences.

5.2 Group similar unique sequences

The core of our analytical procedure is to group similar unique sequences. To this end, the user needs to identify which unique sequences are similar in semantics and time and to construct groups. We facilitate the identification of similar unique sequences by providing a ranking. It is based on a computational similarity measure, which we explain further in Sec. 6. The measure compares the sets of states and the order of states in different categorical sequences. The ranking unburdens the user from the laborious task to identify categorical sequences that are composed of similar sets of states in similar order. Instead, the user can focus on the interpretation and comparison of the semantic meaning.

Our procedure to identify groups of related sequences is iterative. Each iteration results in a novel group of related sequences. At the beginning of each iteration, a unique sequence of interest is identified

from an initial inspection of the semantic and temporal overview provided in step 1 (Fig. 5). A novel group is constructed that initially comprises the unique sequence of interest as a reference sequence. The remaining unique sequences are automatically ranked by the computed similarity of the categorical sequences. When the user lowers the similarity threshold, sequences are automatically added to the group. The similarity threshold is decreased as far as the geoscientist considers the sequences in the group as semantically and temporally similar. This assessment is supported by visual means. Our visualization (Fig. 6) comprises a histogram over time to show the temporal distribution of the group and a depiction of the unique sequences in the group. A brushing and linking concept applied to both views supports the inspection of specific unique sequences.

The computational similarity that is used to rank the unique sequences serves as an indicator for semantic similarity, but the geoscientist eventually assesses which unique sequences are semantically and temporally similar. Hence, the geoscientist can manually adapt the group to remove individual categorical sequences that are not similar to the reference sequence.

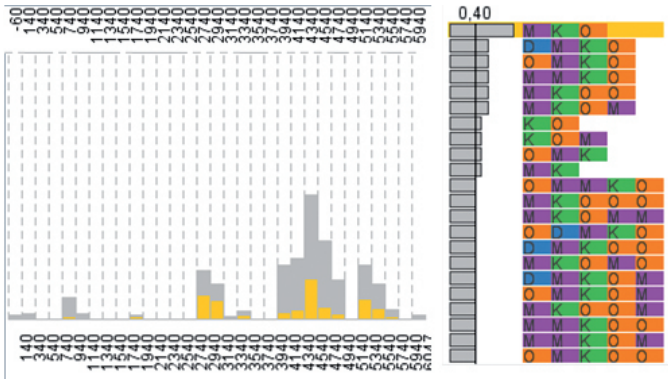


Fig. 6. Step 2: The temporary group of unique sequences is shown by its temporal context and its semantics. The histogram on the left shows the temporal distribution of the sequences in the group. The semantics are depicted by the list of unique sequences on the right. The unique sequences are ranked by their computational similarity to the reference sequence. The small bar charts visualize the computed similarity value, the vertical line depicts the current similarity threshold. Yellow bars in both views highlight the unique sequence of interest.

5.3 Inspect groups of related sequences

During the analytical procedure, the data set is transformed from a time sequence of categorical sequences into groups of categorical sequences. The groups are characterized by unique sequences and their temporal distribution. We visualize the temporal context by individual histograms for each group. The semantics are shown by visualizations of the set of unique sequences for each group, as shown in Fig. 7.

6 Computation of the Similarity Measure for Categorical Sequences

To support the central challenge of identifying similar unique sequences, our visual analysis concepts employs a similarity measure to rank unique sequences according to a sequence of interest. The usefulness of the ranking depends on the quality of the similarity measure. In the following, we discuss an exemplary set of similarity measures and their suitability for our application.

In general, our categorical sequences show variations in length. They comprise a low number of states, but the variability of state sequences is potentially high. Gabadinho et al. [13] investigate common similarity measures for categorical sequences. One important type of measures counts the number of matching attributes between two sequences. They do not directly exploit the sequences of states for determining similarity. An example is the longest common subsequence. In our application, with short and highly variable sequences, this measure is not suitable.

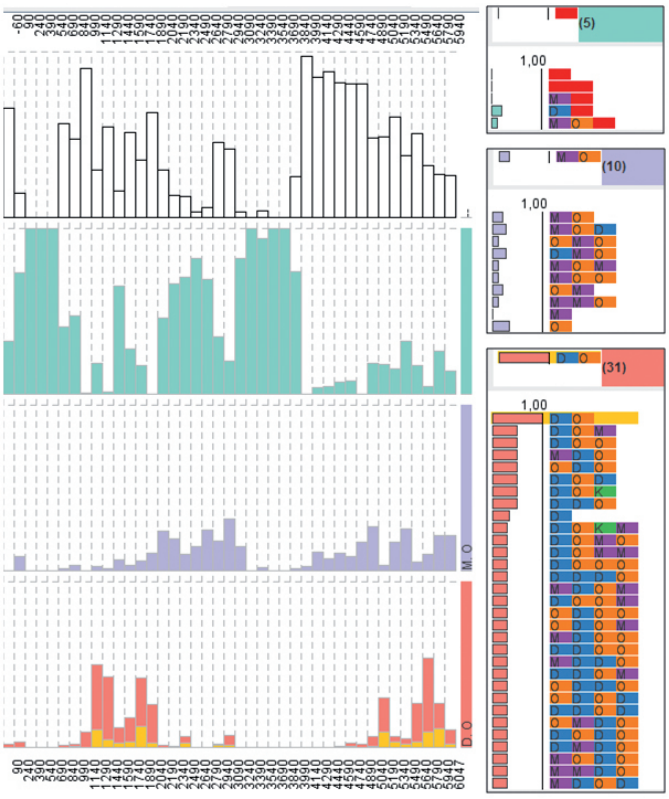


Fig. 7. Step 3: All groups of categorical sequences are explored by the two facets semantics and temporal context. For each group, a histogram depicts the temporal distribution (on the left). The semantics are shown by lists of unique sequences for each group (on the right). Both views are linked by a consistent qualitative color scheme for the groups and by a brushing and linking mechanism to inspect unique sequences.

It ignores parts of the sequences and thereby overlooks similarities. Another important group of measures quantifies the cost of transforming sequences into each other. A common standard is the Levenshtein or edit distance [26]. It counts the minimum number of insertions, deletions, and substitutions that are necessary to transform two sequences into each other.

An alternative approach is to compare feature sets that describe the categorical sequences. We adapted a useful set of features from computational linguistics. Here, sequences of text are often described by the set of subsequences (also denoted as n-grams, with n as the length of the subsequence). As an example, the sequence *DOM* comprises the subsequences *D*, *O*, *M*, *DO*, *OM* and *DOM*. To handle subsequences that occur repeatedly within one categorical sequence, we consider them as separate features in the feature set Ω. The feature set contains redundancies, as shorter subsequences are contained in longer subsequences. But as we do not know in advance which subsequences are important, we treat them all equally. Highly variable data sets show a high number of subsequences in total, but only a sparse set of subsequences occurs in each categorical sequence.

To compare sparse feature sets, the Jaccard-Index [25] is a suitable and established measure. It determines the similarity of the sets Ω_1 and Ω_2 by dividing the size of the intersection by the size of the union.

$$J(\Omega_1,\Omega_2) = \frac{|\Omega_1 \cap \Omega_2|}{|\Omega_1 \cup \Omega_2|} = \frac{|\Omega_1 \cap \Omega_2|}{|\Omega_1| + |\Omega_2| - |\Omega_1 \cap \Omega_2|} \quad (1)$$

The Jaccard-Index computes the fraction of subsequences that are shared by both sequences. The computed similarity values range from $J = 1$ for identical sequences and a value of $J = 0$ for sequences that do not share subsequences. Both the set of states and the sequence of states matter. Sequences that comprise the same set of states in different

order still exhibit some similarity.

To identify a suitable similarity measure, we implemented two exemplary measures: the Levenshtein distance and the Jaccard-Index applied to sets of subsequences as described. We presented the resulting rankings of categorical sequences to the domain expert, who clearly favored the rankings based on the Jaccard-Index. Compared to Levenshtein, it showed a better correspondence to the analysts's notion of similarity in state composition and order.

7 Visual Interface

Our visual interface is composed of two closely linked visualization components: one component for semantics and one component for temporal context. This composition of the visual interface is consistent throughout the analytical procedure. In the following, we discuss the visual design of each component and indicate necessary adaptations to varying requirements during analysis.

7.1 Visualization Component 1: Visualization of semantics

The component bundles visual representations of the data semantics. To assess the semantics, the user inspects and compares unique categorical sequences. Hence, we need to visually represent multiple sequences of categorical states simultaneously. The user should be able to easily identify states within one sequence, to confine different sequences and compare them to each other.

To account for these requirements, we apply a two-dimensional spatial layout as the basic design. One spatial dimension serves to show the states within individual categorical sequences. The second spatial dimension is used to arrange multiple categorical sequences. In our layout, shown in Fig. 8, each row represents one sequence. Multiple sequences are arranged from top to bottom and aligned to the left. With this layout, the user can easily identify individual state sequences (in horizontal direction) and differentiate between multiple sequences as well as compare them (in vertical direction).

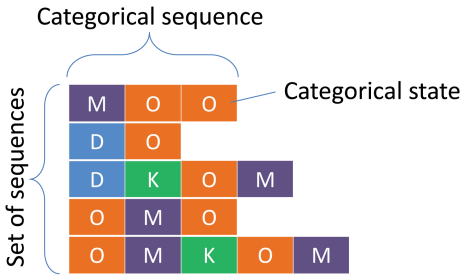


Fig. 8. Basic visualization of categorical sequences. A sequence of states is shown by a sequence of colored blocks with uniform size, from left to right. A block's color represents the categorical state. Multiple sequences are arranged vertically.

Expressive and commonly used visual variables for categorical data are color and shape. For both variables, the number of discernible categories is limited. While combining both variables could yield to a higher number of unique visual primitives, it comes with the cost that the user's efforts for visual inspection are substantially increased [17]. Hence, we use only one visual variable to discern categories, which is color. The qualitative color scheme is adapted from Color Brewer [8]. In our exemplary data sets, the number of categories is rather small and the number of distinctive colors in the color scale sufficient. If more colors are required, we automatically complement the color scale by samples from the CIELAB color space and offer the possibility to adapt colors interactively. In addition to color, categories are made explicit by labels.

We do not use shape to discern states. Instead, all states are mapped to blocks. Blocks indicate the immanent temporal extent of states (in contrast to events). The size of the blocks are uniform, as our data does not deliver sufficient information to explicitly visualize the duration of a state. Aiming for an expressive visualization of state sequences, we avoid visual cues about the duration of states. All states are therefore depicted as uniformly sized blocks. In the result, the visualized length of a sequence corresponds to the number of states in the sequence, no to its time frame (which is known to be one year for all sequences). We also utilize the blocks to group sequences visually in our two-dimensional layout. The blocks' widths are larger than their heights and the margins are larger vertically than horizontally. Thereby, states within one sequence are grouped while different categorical sequences are confined.

We adapt our general visual design to the different sets of unique categorical sequences that appear during the analytical procedure. It results in two views.

View on unique sequences The overview on the semantics shows the automatically extracted unique sequences together with their frequency (left part of Fig. 9). The list of unique sequences is sorted by frequency, with the most frequent on top. Frequency is also explicitly visualized by small bar charts left to the unique sequences. To cope with hundreds of unique sequences, the scrolling mechanism supports a subsequent exploration of the list of unique sequences.

View on groups of similar sequences The view shows the semantics of all groups that have been generated by the user (right part of Fig. 9). All groups are shown by their reference sequence and the list of unique sequences, sorted by the similarity value. The groups are shown below each other, with the most recent group on top. Duplicate group memberships of unique sequences are visually emphasized by color. The computed similarity of unique sequences to the reference sequence plays an important role during the grouping of similar sequences. Hence, the computed similarity values are explicitly shown in a bar chart left to the unique sequences. On top of the bar chart, a vertical line depicts the current similarity threshold.

The two views show two complementary subsets of the data: The first view visualizes all unique sequences that have not been assigned to a group. Sequences that are assigned to a group are shown in the second view. Together, the two views provide access to the complete data set and convey the progress of the analytical procedure, as unique sequences move from the first to the second view. We therefore show both views concurrently beside each other.

As a side effect, the arrangement of views in the component does not need to be adapted during the analytical procedure. This is beneficial during the grouping of similar categorical sequences, which involves repeated consultations of both views.

7.2 Visualization Component 2: Visualization of temporal context

The component subsumes visualization methods that represent temporal context. Our data comprises thousands of time points. That order of magnitude hampers a concurrent in-depth visualization of all time points on average screen resolutions. Useful strategies to derive overviews over thousands of time points depend on the data that needs to be shown.

Considering the groups of unique sequences, the main goal is to show in which time periods the groups appear. The grouping step introduces a novel qualitative variable: Each time point is associated to a group. Temporal histograms are well-suited to show the temporal distribution of groups. We show multiple groups by individual histograms rather than one stacked histogram, even though individual histograms require more screen space. Our main microfacies expert favored individual histograms. They better support inspection and comparison of the temporal distributions of groups as well as the grasping of changes in the histogram during the adaptation of temporary groups. The histograms are placed below each other to facilitate comparison of different groups.

The second data facet that is shown over time is the set of unique sequences. Using histograms for a high number of unique sequences would require the inspection of hundreds of histogram, which is not feasible. To provide a comprehensive overview how unique sequences are associated to time points, we depict the complete time sequence of categorical sequences. An important design decision is the spatial

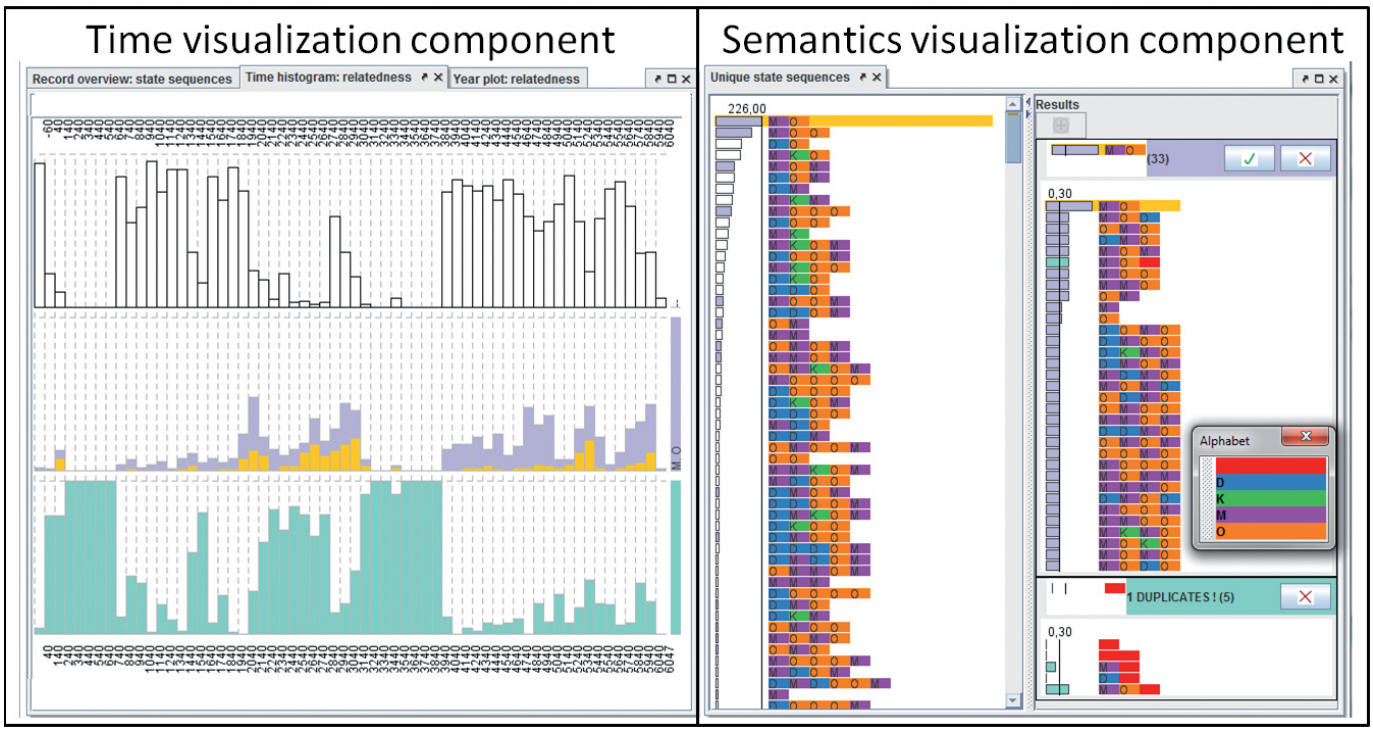


Fig. 9. Our visual interface is composed of two components: A time visualization component on the left and a semantics visualization component on the right. The time visualization shows the temporal distribution of all groups of unique sequences in separate histograms. The component also provides the time overview for the analytical step 1 that is shown in Fig. 5 on the left. The user can switch between both views on demand. The semantics visualization component shows the frequency-sorted set of unique sequences that have not been assigned to a group (left side of component). Beside, the sets of unique sequences that have been grouped are shown in separate lists. The bar chart shows the computational similarity of each unique sequence to the unique sequence of interest (highlighted in yellow). The screenshot was derived during the analysis session that we describe in Sec. 8.1.1. It depicts the process of generating a new group of categorical sequences from the reference sequence MO.

arrangement of the time sequence. E.g., time can be arranged linearly along one spatial axis, or in pixel based or cyclic designs. Our basic arrangement is adapted from the histogram: the time sequence is split into multiple uniform intervals. The uniformly sized time segments are arranged column-wise and the time sequence within an interval is depicted from bottom to top (Fig. 10).

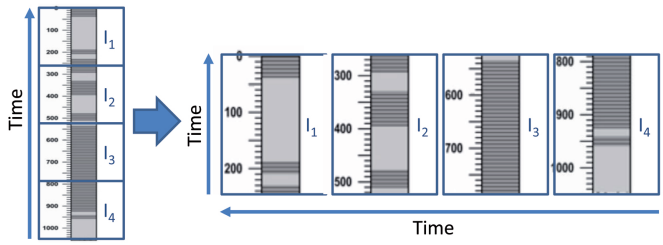


Fig. 10. Spatial arrangement of temporal dimension for the overview on categorical sequences over time: the time sequence is split into intervals that are arranged in columns.

This mapping allows to reuse the visual representation of categorical sequences from the semantics visualization component, which also shows categorical sequences in rows. Within a time interval, time is oriented from bottom to top, which maps with the original orientation of time in the sediment core. By using the same uniform temporal intervals as in the histogram, the general spatial arrangement of time is consistent in both visualizations. In the comprehensive overview, visual representations of categorical sequences are strongly diminished.

To summarize, our component provides three different views.

Time Overview The time overview shows all categorical sequences over time in a highly compact layout, as shown in Fig. 5 on the left. The view scales to larger data sets (in the same order of magnitude) or smaller screen sizes by providing horizontal scrolling.

Detail View The detail view magnifies a temporal subsequence of categorical sequences (Fig. 5).

Histogram The histogram view (left view in Fig. 9) shows the temporal distributions of groups of sequences.

The data shown in the time overview and in the histograms relates to different stages of the analysis procedure. It is therefore sufficient to show one view at a time and offer the user the flexibility to switch between the two views. In addition, the on-demand detail view is shown on top of the visual interface. The selected temporal subsequence is communicated by a black frame in the time overview.

7.3 Visual linking and interaction

The two visualization components are linked via two interaction mechanisms. The first mechanism supports the selection of a unique sequence as part of the brushing and linking concept. The second mechanism involves the definition of a new group of similar sequences.

The selection of a unique sequence is supported and propagated in both visualization components. The realization in the semantic visualization component is straight-forward, as unique sequences are explicitly visualized. In the time visualization component, brushing a time point initializes the selection of the associated unique sequence. The second interaction mechanism supports the definition of a new group of similar sequences. It comprises the generation, adaptation, and storage of a group as well as the handling of duplicate group memberships of unique sequences. Interactive means to derive a group are integrated in the semantics visualization component (Fig. 9). The induced changes are propagated to the other visualization component.

Both visualization components are visually linked. Categorical sequences are consistently mapped to sequences of colored blocks in both components, using a persistent qualitative color scale. In addition, groups of similar sequences are visually linked by the use of a second qualitative color scheme (light qualitative scheme from Color Brewer [8]). One color from the scheme is applied to each group, including the temporary group during the second analytical step. For sequences that have not been assigned to a set of related sequences, the color icon remains white. Further, the unique sequence of interest is visually highlighted by an underlying colored frame (in yellow) in both components.

8 Results

8.1 Use cases

Our tool was evaluated by two experts on microfacies data analysis. They used our tool to perform analyses on two data sets. The data sets originate from two different lakes and exhibit different data characteristics [9, 31].

8.1.1 Use case 1: Investigation of microfacies data with many unique categorical sequences

Initially, our tool provides an overview of the data set (see Fig. 5). The data set spans about 6,000 years. The lengths of categorical sequences vary from 1 to 11 states. Five distinct states of sediment layers are discerned in the data set: M, O, K, D and an unspecified state with unlaminated sediment, which we denote as N. The data set comprises 612 unique sequences. The analyst starts her investigations by exploring the most frequent unique sequences. She performs the same analysis pattern repetitively: First, she selects the next most frequent unique sequence and explores the temporal distribution. Then, she investigates different temporary groups of unique sequences by adapting the similarity threshold.

For the by far most frequent sequence N, she quickly identifies the four unique sequences that also contain the state N and stores them as a group (in turquoise color in Fig. 7). The analyst moves on in the list of frequent unique sequences, to MO. By decreasing the similarity threshold, it becomes apparent that additional sequences predominantly occur in the same temporal periods. In conclusion, 10 categorical sequences are subsumed at the threshold value of 0.33 (light purple in Fig. 7). The next frequent sequence is DO. Also for this group, the temporal periods in which sequences occur do not change for decreased similarity thresholds. That is the basis for defining the group of unique sequences for the reference sequence DO. It comprises 35 unique sequences (light red in Fig. 7).

The temporal distribution of the red group derived from DO is an interesting finding to the analyst (left side in Fig. 7). The sequences in the group appear in two different time periods: from $6,000$ to $5,000$ before present and from $2,000$ to $1,000$ before present. From other studies of landscape development and pollen analysis, the two temporal periods are known to be characterized by different environmental conditions [9]. These differences are attributed to human influence in the latter period. In consequence, different responses of the lake are expected in the two time periods. But our analysis reveals a strong similarity of the responses. This similarity is also confirmed by results from geochemistry analysis of the lake sediment core. From the analysis, geoscientists conclude that the processes in the lake are similar despite differences in environmental conditions.

8.1.2 Use case 2: Investigation of microfacies data with little variability in categorical sequences

The second data set spans a similar time span as the first data set (7,300 years). In contrast to the data set in the first use case, the sequences in the second data set are uniform. All 7.300 show the same categorical sequence CD. In this situation, the analyst makes use of an additional feature of our tool: It supports the redefinition of states based on a user-selected set of state attributes. The analyst discerns the two states C and D by additional attributes. At first, the analyst builds states by the additional attribute $C-Category$. $C-Category$ subdivides the state C into three different states Cf/g, Cg/f and Cf. The data then comprises four uniques sequences (Fig. 11). This allows a plain approach to determine groups: Each unique sequence defines one group. Consequently, four groups are defined.

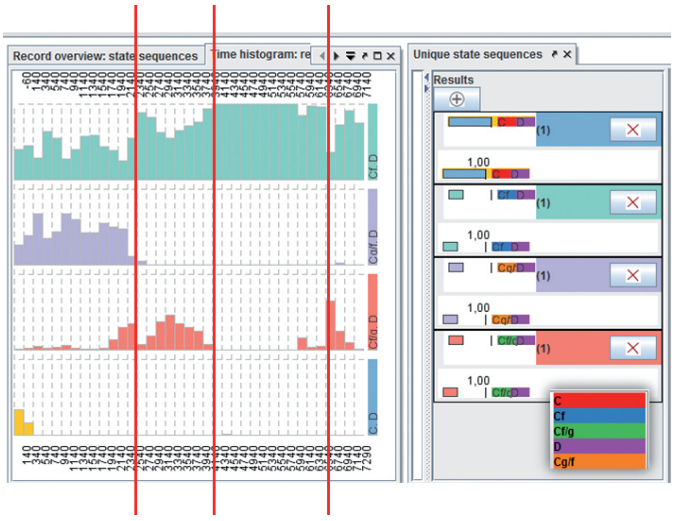


Fig. 11. Analysis results in use case 2. The categorical states are derived from a combination of two qualitative attributes. The time sequence comprises four unique sequences. After defining one group for each unique sequence, the temporal histogram shows three major transitions points: (1) at 2,400, (2) at 4,000, and (3) at 6,400. They confirm findings from other studies [14, 29, 36].

The inspection of the temporal context of the sequences (histograms in Fig. 11 on the left) shows that all sequences are temporally different. Hence, the subdivision of the states by additional attributes is meaningful. The temporal histograms show three major transition points. These transition points are consistent with the finding from other studies [14, 29, 36]. Next, the analyst defines the states from the attributes $D-Category$ and $D-Thickness$. The ordinal attribute $D-Thickness$ results from a conversion of the the numerical thickness into three classes. The set of attributes leads to eight different states and six unique sequences (Fig. 12). The inspection of the temporal similarity of the unique sequences shows similarities among them. The user manually groups the unique sequences that appear in similar time periods. The resulting groups (Fig. 12) are dominantly differentiated by $D-Thickness$ (the red group, turquoise, and purple groups solely differ in values of $D-Thickness$). Hence, $D-Thickness$ is a meaningful indicator of different environmental conditions. $D-Category$ also subdivides the set of unique sequence, but only for smaller values of $D-Thickness$ (orange group), which means it plays a less important role for the differentiation of conditions than $D-Thickness$.

Three temporal transitions are apparent in the histograms which were not known to appear in the data.

Subsequently, the analyst tests other combinations of attributes. They lead to other states and categorical sequences. Their investigation leads to the same transitions as shown in Fig. 12. According to the analyst, the findings can be associated with the known history of climate, biology, and human society.

8.2 Feedback

In the following, feedback from domain experts is stated. First, we describe how we used the expert's comments to evolve our concept during the iterative design process. Several views and functionalities were not accepted by the user. An example was an alternative view to the histogram, which depicts the group labels for every year. It was not considered as useful, because the temporal distribution of groups was difficult to grasp. Further, the domain expert did not utilize interactive means to adjust the ranking function by adapting the underlying feature sets. It was preferred to adapt the groups of sequences directly by removing individual sequences. Expert feedback also initiated a complete

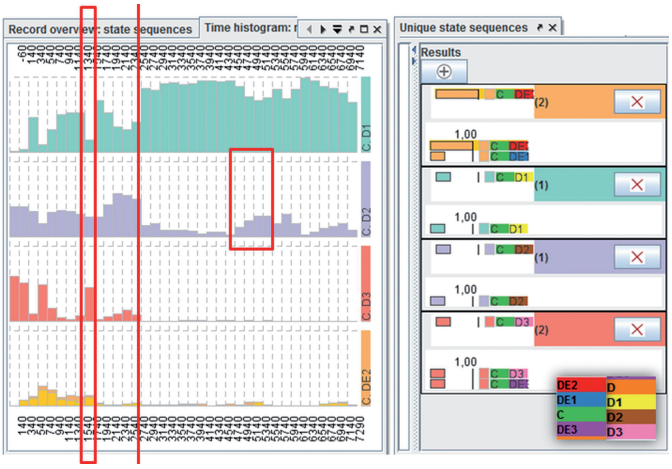


Fig. 12. Analysis results in use case 2 for categorical states from a second attribute combination. One group is defined for each of the six unique sequences in the data set. The temporal histogram reveals novel findings: (1) An abrupt change appears around 1,500 before present. (2) A transition point at 2,000 before present. (3) Gradual changes occur in the time period 4,700 to 5,300 before present.

revision of our visual interface. Initially, we provided dedicated visual interfaces for every analysis step. Switching back and forth between interfaces during the analysis irritated the user. This led to our final view arrangement with two permanently shown major visualization components that can be adapted on demand.

Based on our initial analysis sessions with the final prototype, the two involved geoscientists denoted our approach as the first to enable a comprehensive investigation of the complete time sequence of categorical sequences. During both analysis sessions, time periods of similar climate and environmental conditions known from other studies were confirmed. The ability to generate novel hypotheses was highlighted as a strength of our approach. Also, the reduced effort for investigating hypotheses compared to other methods was emphasized. Our tool unloads the analyst from tedious data processing and allows focusing on domain-specific analytical questions.

Further, we gathered feedback during a demonstration of our tool to eight members of a leading research group in landscape and climate development. During the demonstration, we observed that our tool immediately enforced discussions about dependencies between categorical state sequences and temporal developments. This is a strong indication that our visual interface effectively reveals structures in microfacies data sets. Beyond, our tool was denominated as the first systematic analysis method for microfacies data. It was identified as "a missing bit in our method portfolio", as it allows to fully utilize the information provided by microscopic analysis. Our approach proposes a novel standard for analysis and visual representation of microfacies data, which is generally applicable to data from different lakes.

8.3 Discussion

We conclude from our initial evaluation that our concept is suitable to analyze microfacies data from lake sediment cores. Applied to data set with many unique sequences (use case 1), the geoscientist gains the ability to identify similar categorical sequences. Even though we tailored our approach for highly variable data sets, the evaluation also indicates its usefulness for data sets with little diversity among sequences (use case 2). By redefining states for different sets of attributes, novel findings about temporal developments were derived. The analyst took advantage of the ability to quickly identify the temporal context of unique categorical sequences. The second use case also showed a limitation of our approach. In the investigated data set, semantic relations do not predominantly result from the sequence of states. Instead, they result from the states' multivariate values. Our tool is not designed to handle multivariate relations among states, as we consider all states as categorical. The challenges of defining and considering similarities among states is a current research topic [12].

Our concept has been designed for the specific application of understanding microfacies data. Still, the resulting concept provides a general solution to the problem of finding similar time points in a series of categorical sequences. Our methods do not explicitly incorporate domain knowledge. The automatic extraction of unique sequences, temporal occurrences, and frequency are domain-independent. The similarity measure is general for categorical sequences, but can be changed for other applications. Domain knowledge is introduced into the analytical process by human assessment, it is not formalized within the methods. In this regard, our approach is adaptable to other application fields with similar data and similar tasks.

The scalability of our approach is affected by number of states, lengths of sequences, number of unique sequences, and number of time points. Our concept scales well to domain-specific data sets. Considering the visual design, it also adapts to longer sequences, more unique sequences and more time points within the same order of magnitude. A crucial aspect for scalability of the visual design is the number of states. More states significantly affect the user's ability to discern sequences of states. We estimate an upper limit of 10 to 12 categories. This limitation is not specific to our approach, but presents a general analysis challenge for categorical data.

Considering the scalability of computational methods, the most relevant operation is the ranking of sequences. It is performed numerous times during the analysis. The ranking involves two steps: the feature set of every unique sequence is generated and the similarity measure is applied. We compute the feature sets only once, at the beginning of the analysis. During analysis, the similarity values are computed very fast. The Jaccard-Index requires one set operation and few algorithmic operators per unique sequence. The computational effort increases with the number of unique sequences and the sequence lengths, which directly affect the sizes of the feature sets. In several experiments, we doubled and quadrupled the feature sets and the unique sequences compared to the data set in Sec. 8.1.1. While the initial generation of the feature sets required more time (in the range of seconds), we did not observe that the computation time for the ranking of sequences increased significantly.

9 Conclusion

We have introduced a visual analysis concept that supports the analysis of microfacies data from sediment cores. Our approach enables analytical investigations that were not carried out before due to high analytical efforts. We conclude from our use cases that our prototype facilitates the identification of similar sequences. The analytical results reveal interesting findings, whose relevance has been confirmed in other geoscientific studies. According to the domain experts, our approach defines common standards for the analysis of microscopic data from sediment cores of different lakes. Thereby, we close an important methodological gap in the application domain, as our concept promotes the utilization of the rich information provided by microscopic analyses from sediment cores.

A major benefit of using a visual analytics approach is that it allows geoscientists maintaining their domain-specific perspective during the analysis. With our visual interface, geoscientists focus on solving their domain-specific analytical problem instead of learning how to deploy computational analysis methods.

Acknowledgments

This study is a contribution to the Virtual Institute of Integrated Climate and Landscape Evolution Analysis (ICLEA) of the Helmholtz Association (grant number VH-VI-415).

We would like to thank all involved scientists in GFZ-section 5.2., particularly Achim Brauer, Florian Ott and Ulrike Kienel. Further, we thank Patrick Köthur, Bin Yang, Daniela Rabe and Johannes Hofmann for their valuable comments and support in preparing the paper and additional materials.

REFERENCES

[1] C. C. Aggarwal. *Data Mining: The Textbook*. Springer Publishing Company, 2015.

[2] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2011.

[3] D. Albers, C. Dewey, and M. Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE transactions on visualization and computer graphics*, 17(12):2392–2401, 2011.

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pp. 217–235. Springer-Verlag, London, UK, 1999.

[5] A. Brauer. Annually laminated lake sediments and their palaeoclimatic relevance. In *The Climate in Historical Times*, pp. 109–127. Springer, 2004.

[6] A. Brauer, P. Dulski, C. Mangili, J. Mingram, and J. Liu. The potential of varves in high-resolution paleolimnological studies. *PAGES news*, 17(3):96–98, 2009.

[7] A. Brauer, C. Mangili, A. Moscariello, and A. Witt. Palaeoclimatic implications from micro-facies data of a 5900 varve time series from the Piànico interglacial sediment record, southern Alps. *Palaeogeography, Palaeoclimatology, Palaeoecology*, 259(2-3):121–135, 2008. doi: 10.1016/j.palaeo.2007.10.003

[8] C. A. Brewer, G. W. Hatchard, and M. A. Harrower. Colorbrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(28):5–32, 2003.

[9] N. Dräger, M. Theuerkauf, K. Szeroczynska, S. Wulf, R. Tjallingii, B. Plessen, U. Kienel, and A. Brauer. Varve microfacies and varve preservation record of climate change and human impact for the last 6000 years at Lake Tiefer See (NE Germany). *The Holocene*, 27(3):450–464, 2017. doi: 10.1177/0959683616660173

[10] F. Du, B. Shneiderman, C. Plaisant, S. Malik, and A. Perer. Coping with volume and variety in temporal event sequences: Strategies for sharpening analytic focus. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016. doi: 10.1109/TVCG.2016.2539960

[11] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid. Periodicity detection in time series databases. *IEEE Transactions on Knowledge and Data Engineering*, 17(7):875–887, July 2005. doi: 10.1109/TKDE.2005.114

[12] C. H. Elzinga and M. Studer. Spell sequences, state proximities, and distance metrics. *Sociological Methods & Research*, 44(1):3–47, 2015.

[13] A. Gabadinho, G. Ritschard, N. Müller, and M. Studer. Analyzing and visualizing state sequences in r with traminer. *Journal of Statistical Software*, 40(1):1–37, 2011. doi: 10.18637/jss.v040.i04

[14] J. N. Haas, I. Richoz, W. Tinner, and L. Wick. Synchronous holocene climatic oscillations recorded on the swiss plateau and at timberline in the alps. *The Holocene*, 8(3):301–309, 1998. doi: 10.1191/095968398675491173

[15] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd revised edition ed., 2011.

[16] M. Hao, U. Dayal, D. Keim, and T. Schreck. Multi-resolution techniques for visual exploration of large time-series data. In *Proceedings of the 9th Joint Eurographics / IEEE VGTC Conference on Visualization*, EUROVIS'07, pp. 27–34. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 2007. doi: 10.2312/VisSym/EuroVis07/027-034

[17] C. Healey and J. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7):1170–1188, July 2012. doi: 10.1109/TVCG.2011.127

[18] A. Hinneburg, C. C. Aggarwal, and D. A. Keim. What is the nearest neighbor in high dimensional spaces? In *VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases*, pp. 506–515. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

[19] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. Dna visual and analytic data mining. In *Proceedings of the 8th conference on Visualization '97*, pp. 437–ff. IEEE Computer Society Press, Los Alamitos, CA, USA, 1997.

[20] W. Javed and N. Elmqvist. Stack zooming for multifocus interaction in skewed-aspect visual spaces. *IEEE Transactions on Visualization and Computer Graphics*, 19(8):1362–1374, Aug 2013. doi: 10.1109/TVCG.2012.323

[21] E. Kandogan. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. *In Proceedings of the IEEE Information Visualization Symposium*, 2000.

[22] D. A. Keim, M. C. Hao, U. Dayal, and M. Hsu. Pixel bar charts: A visualization technique for very large multi-attribute data sets. *Information Visualization*, 1:20–34, 2002.

[23] K.R.Gabriel. The biplot graphic display of matrices with application to principal component analysis 1. *Biometrika*, 58(3):453–467, 1971.

[24] D. J. Lehmann and H. Theisel. Orthographic star coordinates. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2615–2624, 2013.

[25] M. Levandowsky and D. Winter. Distance between sets. *Nature*, 234:34–35, 1971.

[26] V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707 – 710, 1966.

[27] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *ACM Intelligent User Interfaces (IUI) 2015*, 2015.

[28] C. Martin-Puertas, K. Matthes, A. Brauer, R. Muscheler, F. Hansen, C. Petrick, A. Aldahan, G. Possnert, and B. van Geel. Regional atmospheric circulation shifts induced by a grand solar minimum. *Nature Geoscience*, 5(6):397–401, may 2012. doi: 10.1038/ngeo1460

[29] C. Martin-Puertas, K. Matthes, A. Brauer, R. Muscheler, F. Hansen, C. Petrick, A. Aldahan, G. Possnert, and B. van Geel. Regional atmospheric circulation shifts induced by a grand solar minimum. *Nature Geoscience*, 5:397–401, 2012.

[30] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236, 2013.

[31] F. Ott, A. Brauer, M. Słowiński, S. Wulf, V. Putyrskaya, B. Plessen, and M. Błaszkiewicz. Varved sediments from Lake Czechowskie (Poland) reveal gradual increase in Atlantic influence during the Holocene. *Geophysical Research Abstracts*, 17:EGU2015–308, 2015.

[32] J. G. Paiva, L. Florian, H. Pedrini, G. Telles, and R. Minghim. Improved similarity trees and their application to visual data classification. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2459–2468, 2011.

[33] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: using visualization to enhance navigation and analysis of patient records. In *Proceedings of the AMIA Symposium*, p. 76. American Medical Informatics Association, 1998.

[34] B. Rieck and H. Leitte. Exploring and comparing clusterings of multivariate data sets using persistent homology. *Computer Graphics Forum*, 35(3):81–90, 2016. doi: 10.1111/cgf.12884

[35] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 31:1335–1344, 2012.

[36] T. M. Shanahan, N. P. McKay, K. A. Hughen, J. T. Overpeck, B. Otto-Bliesner, C. W. Heil, J. King, C. A. Scholz, and J. Peck. The time-transgressive termination of the african humid period. *Nature Geoscience*, 8:140–144, 2015.

[37] S. F. Silva and T. Catarci. Visualization of linear time-oriented data: A survey. In *Proceedings of the First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1 - Volume 1*, WISE '00, pp. 310–. IEEE Computer Society, Washington, DC, USA, 2000.

[38] K. Vrotsou, J. Johansson, and M. Cooper. Activitree: Interactive visual exploration of sequences in event-based data using graph similarity. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):945–952, 2009.

[39] K. Vrotsou, A. Ynnerman, and M. Cooper. Are we what we do? exploring group behaviour through user-defined event-sequence similarity. *Information Visualization*, pp. 232–247, 2013.

[40] J. Walker, R. Borgo, and M. W. Jones. Timenotes: A study on effective chart visualization and interaction techniques for time-series data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):549–558, Jan 2016. doi: 10.1109/TVCG.2015.2467751

[41] J. S. Walker, M. W. Jones, R. S. Laramee, O. R. Bidder, H. J. Williams, R. Scott, E. L. C. Shepard, and R. P. Wilson. Timeclassifier: a visual analytic system for the classification of multi-dimensional time series data. *The Visual Computer*, 31(6-8):1067–1078, 2015. doi: 10.1007/s00371-015-1112-0

[42] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1747–1756. ACM, 2011.

[43] K. Wongsuphasawat and B. Shneiderman. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 27–34. IEEE, 2009.

[44] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '00, pp. 275–279. ACM, New York, NY, USA, 2000. doi: 10.1145/347090.347150

[45] X. Zhang, F. Pan, and W. Wang. Finding high-order correlations in high-dimensional biological data. In *Link Mining: Models, Algorithms and Applications (Eds. Yu, Han, and Faloutsos)*, pp. 505–534, 2010.

[46] Y. Zhang, W. Luo, E. A. Mack, and R. Maciejewski. Visualizing the impact of geographical variations on multivariate clustering. *Computer Graphics Forum (In Proc. EuroVis)*, 2016.