SMART Series: Sketch-based Matching through Approximated Ratios in Time Series

Prithiviraj K. Muthumanickam Katerina Vrotsou Matthew Cooper Jimmy Johansson*

Linköping University, Sweden

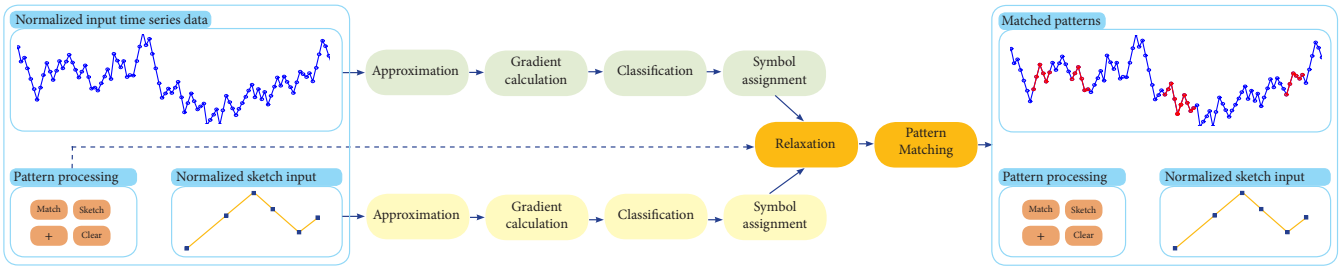

Figure 1: Schematic representation of the SMART Series workflow. A pattern of interest sketched by the user is subsequently matched to the time-series data. Efficient approximation, classification and symbol assignment, based on ratios, enables real-time pattern searching within very large time-series. The steps depicted with green boxes in the figure are executed only when a new input time series is loaded, yellow boxes only when a new sketch is entered, and orange boxes only when a new tolerance relaxation factor is applied for the pattern matching.

Abstract

Searching for all possible patterns in a time series graph is a computationally complex problem. User-sketched pattern matching is an effective semi-automatic approach to address this problem but the search space is still very large the accuracy of the pattern search must be considered. Our method, SMART series, uses a ratio-based approximation of the raw time series and transforms it into a symbolic representation. The user can then draw patterns of interest in a separate sketching space and search, in real time, for matches within the symbolic space. Accuracy relaxation in the matching is provided through a further reduction of the symbolic space.

Index Terms: H.3.3 [Information Systems]: Information search and retrieval; H.5.2 [Information interfaces and presentation]: GUI; I.3.6 [Computing Methodologies]: Interaction Techniques;

1 Introduction

Improved sensor networks and storage facilities are resulting in very large time series data being generated within many application domains. Temporal data are commonly represented as time series graphs but, for long series, such representations become cluttered thereby reducing the visibility of interesting characteristics among them. Identifying specific features such as recurring patterns, outliers and anomalous trends are therefore becoming essential analysis tasks. Since, looking for all possible patterns in real time is a computationally complex problem, numerous approaches have been adopted for identifying patterns of interest within time series data. Due to space restrictions we only refer to some of the most recent and closely related work. Some methods achieve this through dimensionality reduction [3] by segmenting the data based on user specified segment length, assigning a symbol to each segment and using a sliding window of user specified length to generate a symbolic approximation. Also, grammar analysis is performed on approximated data to find possible sets of patterns [4]. These methods suffer from a high level of approximation applied at an overall level, which may inhibit pattern search.

Another interesting approach for the identification of time series behaviour is the use of semi-automatic approaches such as user-sketched pattern matching in time series [2]. This can be performed by using rubber-band rectangles or user sketches that are drawn directly on the raw time series graph [1]. These methods suffer from a lack of flexibility and tedious panning through long time series to visually select patterns of interest. The work of Gregory and Shneiderman [2] identifies certain basic shapes in time series such as spikes, sinks, rise, drop, plateau, valley and gaps. For performing a pattern search, users then have to select any of these shapes, along with the number of data points that constitute the shape. Searching for combinations of such shapes is, however, not discussed in this work and additional user input adds to the cognitive load, making it difficult to search for longer patterns. In order to overcome these limitations we propose a hybrid approach that has a reduced amount of user interaction, provides additional flexibility for sketching of patterns and searching for matches in real time. The precision of the matches can be easily controlled by the user. The main contributions of this work are:

- Ratio approximation of time series data, followed by symbolic representation.
- User-sketched pattern matching with a 'one-click' option for precision relaxation without the need for extensive user input.
- Pattern matching and relaxation efficiently performed at the symbolic level rather than within the raw data.

2 Ratio Approximated Representation (RARE)

The application described in this poster allows a user to search for any pattern of interest by sketching an approximation of this pattern in a graphical user interface. To achieve this we consider the time series data as a combination of basic elementary shapes that are positioned across different amplitudes (see figure 2). The core idea behind our approach is that if we break down the entire time

*e-mail:{prithiviraj.muthumanickam,katerina.vrotsou,matthew.cooper, jimmy.johansson}@liu.se

$S_j = 0, S_{j+1} = 0$	$S_j = 0, S_{j+1} > 0$ $S_j > 0, S_{j+1} = 0$	$S_j = 0, S_{j+1} < 0$ $S_j < 0, S_{j+1} = 0$	$S_j > 0, S_{j+1} < 0$ $\|S_j\| > \|S_{j+1}\|$	$S_j > 0, S_{j+1} < 0$ $\|S_j\| < \|S_{j+1}\|$	$S_j < 0, S_{j+1} > 0$ $\|S_j\| > \|S_{j+1}\|$	$S_j < 0, S_{j+1} > 0$ $\|S_j\| < \|S_{j+1}\|$	$S_j > 0, S_{j+1} > 0$ $S_j < 0, S_{j+1} < 0$	$S_j > 0, S_{j+1} < 0$ $\|S_j\| = \|S_{j+1}\|$	$S_j < 0, S_{j+1} > 0$ $\|S_j\| = \|S_{j+1}\|$
Case1	Case2	Case3	Case4	Case5	Case6	Case7	Case8	Case9	Case10

Figure 2: The set of unique cases identified as elementary shapes within a time series.

series into these elementary shapes, it helps us to approximate similar shapes using only a small set of symbols. We now describe the process of our approach as illustrated in figure 1.

Normalization and approximation. We first normalize the input time series data and the user-sketched pattern. An initial approximation is performed on each, where consecutive positive and negative gradients of different lengths are replaced with gradients of equal length. Dissimilarities in pattern matches that arise from amplitude, translation are handled using ratios of gradients, while the scaling problem is addressed by our approximation step.

Gradient calculation. We calculate the gradients s_j and s_{j+1} from time series points t_j, t_{j+1}, t_{j+2} by sliding a window with a length of 3 time points across the time series.

Classification. The obtained gradients are grouped into a finite set of cases making up the basic building blocks of the time series, as shown in figure 2. We perform this classification in order to produce building blocks that can be represented as a ratio value.

Symbol assignment. We compute the ratios of all gradients, s_j and s_{j+1}, falling under each particular case. If s_j or s_{j+1} is zero then we consider the non-zero gradient to be our current ratio. Similar gradient pairs, having close ratio values, are grouped together and represented as a single symbol. Figure 3 portrays the complete process where all the ratio values falling under one of the cases are first sorted and a uniform binning is applied individually to each of the cases except 1 and 8, as their consecutive combinations can be represented as a single symbol. When the user-sketched input pattern is analysed, the computed ratio values are mapped to the closest of the binned ratios of the input time series and converted to their symbolic values.

Pattern matching. We initiate a simple linear time substring search algorithm with the symbolic representations of the time series and user-sketched pattern as input. The resulting index list, containing all occurrences of the sketched pattern, is used to highlight them in the raw time series. The proposed algorithm can be extended to accomodate absolute value matching where the matched patterns can be restricted within an amplitude range.

Relaxation. The matching algorithm performs a closest match for a user-sketched pattern. Further relaxation of the tolerance in the matching can be performed through the application user interface. This relaxation is performed within the symbol space rather than on the raw time series data meaning no re-analysis of the data is required. This process scales linearly with the size of the symbolic representation of the time series. For example, if the binning space of case 5 is represented through symbols *M,N,O,P*, the first level of relaxation will replace all occurrences of *'M'* with *'N'*. Further relaxation will replace all occurrences of *'N'* with *'O'* and so on.

3 Conclusion and Future Work

We have introduced our initial work on an interactive sketch-based pattern search algorithm that works in real time with significant level of accuracy and without complex user interaction. For a time series of length *10k*, the time taken for symbolic approximation, relaxation and pattern matching is of the order of milliseconds. This is achieved through careful approximation by breaking the raw data into a set of basic shapes and computing their ratios, whereby removing the amplitude information associated with them. Such

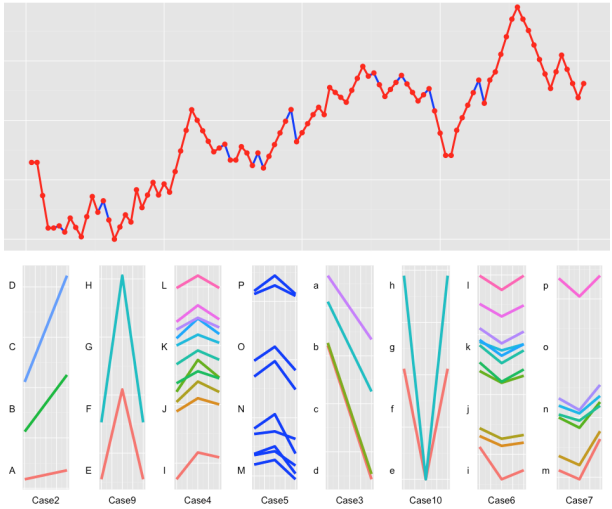

Figure 3: **Top:** Time series after our approximation step. **Bottom:** Adjacent gradients that fall under any of the cases of Figure 2 are mapped to their corresponding columns. Similar ratio values in each column are binned accordingly and represented with the same color. For example, Case 5 with its ratio values and their corresponding gradients are highlighted in blue. The final symbolic approximation of the raw time series data is *dxdAMnKxnyKkOxjyKmJkyLlKnyyy.....xkyMxxbDyyyyOhyyyyyHxxxxxpyLxxl.*

a representation aids us in compressing the symbolic data due to the inherent nature of occurrence of similar patterns. In future we plan to explore advanced binning algorithms, along with prefix tree based approaches to search for similar patterns of different lengths. Grammar based approaches and other text compression algorithms will also be experimented with in order to further compress the approximated symbolic data.

Acknowledgements

This work is funded by the Swedish Research Council, grant number 2013-4939.

References

[1] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proceedings of Conference on Visualization and Data Analysis*, pages 175–186. SPIE, 2005.

[2] M. Gregory and B. Shneiderman. Shape identification in temporal data sets. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 305–321. Springer, 2012.

[3] S. Malinowski, T. Guyet, R. Quiniou, and R. Tavenard. 1d-sax: A novel symbolic representation for time series. In *Advances in Intelligent Data Analysis XII*, pages 273–284. Springer, 2013.

[4] P. Senin, J. Lin, X. Wang, T. Oates, S. Gandhi, A. P. Boedihardjo, C. Chen, S. Frankenstein, and M. Lerner. Grammarviz 2.0: a tool for grammar-based pattern discovery in time series. In *Machine Learning and Knowledge Discovery in Databases*, pages 468–472. Springer, 2014.