Interactive Visual Sequence Mining based on Pattern-Growth

Katerina Vrotsou*
Linköping University, Sweden

Aida Nordman†
Linköping University, Sweden

Abstract

Sequential pattern mining aims to discover valuable patterns from datasets and has a vast number of applications in various fields. Due to the combinatorial nature of the problem, the existing algorithms tend to output long lists of patterns that often suffer from a lack of focus from the user perspective. Our aim is to tackle this problem by combining interactive visualization techniques with sequential pattern mining to create a "transparent box" execution model for existing algorithms. This paper describes our first step in this direction and gives an overview of a system that allows the user to guide the execution of a pattern-growth algorithm at suitable points, through a powerful visual interface.

Index Terms: H.5.2: User Interfaces; H.2.8: Data mining

1 Introduction

Sequential pattern mining is an important field of research addressing the problem of detecting interesting subsequences of events as patterns [1]. One of the main issues with sequential mining algorithms is that they extract too many patterns, many of which may, in fact, be irrelevant. To address this problem, imposing constraints on the sequential patterns has been researched. Constraints are, however, commonly imposed at the start of the algorithm excluding the expert user from the mining process. An interesting problem is how to involve the expert in this process by embedding interactivity deeper in it. To this end, we aim to investigate the possibility of breaking down existing algorithms into incremental steps making it possible to stop the mining process, displaying the current status and allowing a user to intervene by imposing constraints.

This paper presents our first step in this direction and describes briefly a system based on the *PrefixSpan* algorithm [2]. The novelty of the proposed approach is twofold. Firstly, the user can visualize in a stepwise manner the frequent subsequences being built (Fig. 1). Secondly, the approach opens for the possibility of setting separate constraints on different patterns being created. In contrast to other systems [3], the user can refine locally the minimum support parameter for different subsequence patterns or simply stop the mining process of certain subsequences, while allowing other (shorter) frequent subsequences to grow further.

2 Pattern-Growth based Approach

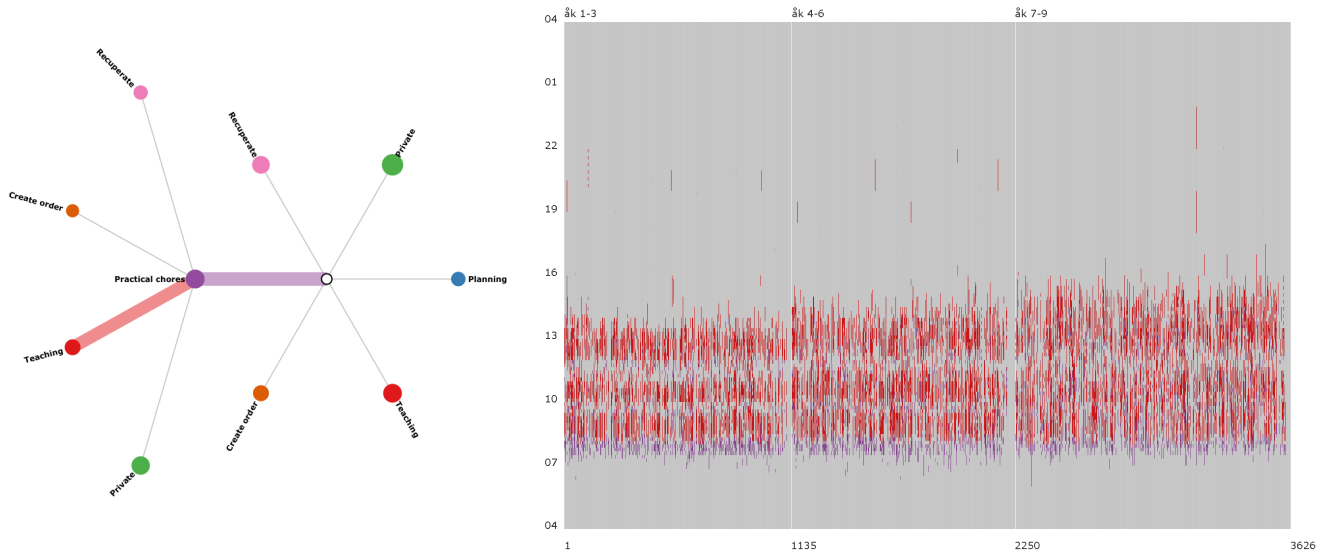
We base our model on the pattern growth methodology which adopts a projection-based divide-and-conquer strategy to frequent sequence pattern mining [2]. A pattern is frequent with respect to a predefined support threshold *minsup*, $0 \leq minsup \leq |\mathrm{DB}|$, where DB is the database of sequences. The underlying algorithm of our system is *PrefixSpan*, which can be summarized as follows. The sequential patterns of length $l \geq 1$ form the prefixes of the frequent patterns of length $l+1$. For each prefix sequence p, *PrefixSpan* builds a projected database consisting of the suffixes of the sequences with prefix p. Each projected databased is then locally mined for its frequent events. Each of these events is then appended to the prefix p to form a longer sequential pattern of length $l+1$.

Figure 1: Examples of the tree representation composed of sequence patterns grown to various lengths.

The pattern growth methodology underlying the *PrefixSpan* algorithm has several advantages. Firstly, no candidate sequences need to be generated reducing in this way the search space. Secondly, patterns are grown incrementally by length, providing in this way points in the execution of the algorithm where it can be paused. At these points, a visual interface can then display a meaningful layout of the patterns that the algorithm has uncovered so far (Fig. 2). Thirdly, growing a pattern p of length $l > 0$ to a pattern of length $l+1$ is done by mining locally a separate database $\mathrm{DB}|_p$, called projected database [2]. This is a key feature for creating a flexible user interface in the proposed system, since it opens for the possibility of setting separate constraints on the patterns being created. For instance, the user can modify the *minsup* parameter specifically for a pattern prefixed by $p = \texttt{event}_1 \rightarrow \texttt{event}_2$ or decide that any pattern prefixed by p is uninteresting and, therefore, stop growing it. Another possibility is that if additional information exists about the events, e.g. time of year, then the user could set that events added to pattern p should occur during summer months.

*e-mail:katerina.vrotsou@liu.se
†e-mail:aida.nordman@liu.se

IEEE Symposium on Visual Analytics Science and Technology 2014
November 9-14, Paris, France
978-1-4799-6227-3/14/$31.00 ©2014 IEEE



(a) Interactive tree representation.

(b) Representation of mined patterns across the data.

Figure 2: The proposed interactive sequence mining system composed of two coordinated views: (a) a tree representation for systematically growing patterns of interest, and (b) an accompanying view for displaying the distribution of the grown patterns across the data. The frequent sequence pattern `practical chores` → `teaching`, identified in the teacher time-use dataset, is highlighted in both representations.

3 Interactive Visual Mining Interface

The proposed system extends the VISUAL-TimePAcTS tool [5] by incorporating the interactive pattern growth methodology based on *PrefixSpan*. The system is composed of two linked views for interactively steering the mining process and inspecting its stepwise results (Fig. 2). A tree representation, based on the Reingold-Tilford algorithm [4], is used for interacting with the *PrefixSpan* algorithm and displaying the patterns as they are mined (Fig. 2(a)). By clicking on the nodes of the tree representation, the user decides which patterns to grow and thus dynamically controls the execution of the algorithm. Sequence patterns are grown outward starting from the root of the tree. The size of the nodes reflects the support of each pattern, decreasing as the tree expands. As patterns are grown their distribution across the data is displayed in an accompanying view (Fig. 2(b)). In this view the event sequences composing the dataset are displayed as vertical bars consisting of the different events and ordered along the x-axis. In both representations of the system colour reflects the event type category.

An initial support threshold is chosen by the user. At each pattern growth level the user has the possibility to right-click on a node and set a new local threshold for the following database projection. Apart from dynamically controlling the mining process a user can also choose to grow all branches of the tree directly, i.e. allow the algorithm to run until completion. The user can then explore the distribution of the patterns across the data by making selections within the tree and inspecting them in the accompanying view.

A dataset collected by the Swedish National Agency for Education for studying the time-use of primary and secondary school teachers is used in this paper for exemplifying our system[1]. The dataset is composed of 3626 teachers' activity sequences. Figure 2 shows an example of a step in the interactive mining process. The user has chosen to grow patterns starting with teachers performing *practical chores*, such as preparing solutions for an experiment, by clicking the `practical chores` node in the tree (Fig. 2(a)). The size of the nodes grown from `practical chores` indicates that one of the most frequently occurring activity following *practical chores* is *teaching*. Selecting this sequence in the tree displays its distribution across the data in the accompanying view (Fig. 2(b)).

[1] http://www.skolverket.se/publikationer?id=3001

In this view, the teachers are displayed along the x-axis grouped by grade; grade 1-3 is displayed in the leftmost group, grade 4-6 in the middle, and grade 7-9 in the rightmost. Time is displayed on the y-axis revealing the activities performed by teachers over a day. Highlighting the sequence pattern in this view shows that the pattern is evenly distributed among teachers of all grade groups and it is mostly performed in the morning before starting the classes.

4 Conclusion and Future Work

This paper presents a system whose interface allows the user to stepwise visualize the frequent sequences being mined and define separate constraints for different patterns being built. In this way, the user can guide the execution of the underlying mining algorithm at suitable points. The combination of the two views provides additional context to the mining process by revealing how the patterns appear in the data and can therefore provide guidance to the user.

We envision an approach that actively embeds interaction within the mining algorithm. In this respect, we plan to include context in the mining process by allowing the user to incorporate background knowledge even after the mining has started. Additional visual cues will also be investigated to reveal information about the current mining status, giving in this way 'direction' to the user performing the search and reinforcing the feedback loop. Finally, the integration of the mining process with ontologies of events will be investigated in order to allow the user to set locally different levels of detail for each of the patterns being mined.

References

[1] C. H. Mooney and J. F. Roddick. Sequential Pattern Mining - Approaches and Algorithms. *ACM Computing Surveys*, 45(2), 2013.

[2] J. Pei, J. Han, B. Mortazavi-asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu. Mining Sequential Patterns by Pattern-Growth : The PrefixSpan Approach. *IEEE Trans on Knowledge and Data Engineering*, 16(11):1424–1440, 2004.

[3] A. Perer and F. Wang. Frequence : Interactive Mining and Visualization of Temporal Frequent Event Sequences. In *International Conference on Intelligent User Interfaces*, pages 153–162, Haifa, Israel, 2014. ACM.

[4] E. Reingold and J. Tilford. Tidier Drawings of Trees. *IEEE Transactions on Software Engineering*, SE-7(2):223–228, Mar. 1981.

[5] K. Vrotsou. *Everyday mining: Exploring sequences in event-based data*. Doctoral thesis, Linköping Studies in Science and Technology. Dissertations No. 1331. Linköping University, 2010.