The Critical Role of Data Mining for Analyzing Real-World Event Sequences

Adam Perer, Bum Chul Kwon, and Janu Verma

1 Introduction

There are many existing techniques for visualizing temporal data, but the complexity of real-world temporal data may undermine certain approaches. Many existing techniques follow consistent and helpful conventions of mapping time to the horizontal axis, aligning multiple timelines along the vertical axis, and assign various event types to categorical visual encodings. A variety of techniques, such as LifeLines [7] and EventFlow [3], have shown great promise in aiding the analysis of temporal patterns. However, as the volume of temporal data increases, along with the data's inherent noise, we argue that visualization and interaction techniques alone may not empower analysts sufficiently. In particular, the diversity of event types, the presence of both relevant and irrelevant event types mixed together, and the lack of strict orderliness in real world data pose many challenges. To this effect, we argue that integrating data mining techniques is critical for making sense out of real-world temporal data for many domains. While automated data mining techniques can't be a substitute for user-centered exploration, data mining techniques can empower users in critical ways.

Visualizing entire event sequences because impossible as the volume of data grows. Simplifying raw event sequences using filtering and aggregation is promising [3], but this may create a reliance of domain expertise. Often, it may be hard to see patterns due to unimportant events distributed through the sequences. In addition, temporal visualization techniques often do not scale to thousands of different event types. However, if we can utilize data mining techniques that effectively unearth statistically significant patterns, simpler and more concise visualizations can be used for communication by focusing on these patterns.

2 Motivating Example

Let's consider an actual case study to illustrate the complexity of real-world temporal event data. Alongside clinical researchers at a healthcare institution, we are interested in analyzing the temporal patterns of patients. We believe that time-stamped patient records, ubiquitous in clinical databases such as electronic medical records, reflect the nature of patient care. In order to assess the feasibility of such analysis, let's begin with only a small subset of their data.

Our investigation begins with a carefully constructed cohort of about 35,000 patients (which is comparatively small to the millions of patients in the full patient databases we hope to analyze). Of these patients, some are cases diagnosed with congestive heart failure (CHF) and the remaining are matched controls who do not have the disease but have matching demographics. The goal is to understand if the patients with CHF have any distinct patterns of disease progression compared to those without the disease. For this initial investigation, we consider only 1 year of data (even though patients may have more than 10 years of data). For patients with CHF, we consider the year leading up to their diagnoses of CHF. For the matched controls, we consider their most recent year. If we only look at their diagnostic codes over the course of the year, we will end up with over 900,000 events that can be categorized into more than 5000 different types of diagnoses [1]. Keep in mind these are only the diagnostic events, so we are ignoring the thousands of unique procedures, treatments, and lab tests that may also have occurred during this year. The event sequences of the patients are very diverse, as some patients have as many over 700 diagnoses during this year, whereas others have none (the average is about 25).

Although this case study only involves the analysis of a *subset* of events types for a *subset* of time for a *subset* of patients, the amount of data is already overwhelming for most current visualization techniques and would rely on advanced filtering or aggregation strategies by users.

Our experience with existing tools like CareFlow [4] and EventFlow [3] often implicitly require users to choose important event types in order to focus the analysis among the countless event types that appear to less relevant to user goals. Though these tools are often technically capable of handling thousands of event types, the visualization metaphors may break down if users were to load so much diverse data, with a plethora of nodes and colors representing all of the event types. Curation is a powerful and necessary feature for users to focus their attention, but it is also possible that unexpected event types might be overlooked and serendipitous discoveries are limited.

3 Is Data Mining the answer?

There is a recent trend to integrate data mining techniques with visualizations in order to reduce the visualized data to only salient patterns. For instance, Frequence [5], Care Pathway Explorer [6], and TimeStitch [8] all use frequent sequence mining techniques to find the most frequent sequences of events.

Many data mining algorithms are not fully automated. For instance, our implementation of a popular and fast mining algorithm, SPAM [1] requires two parameters: *support*, which is the minimum number of patients that must have an event sequence to quality as a pattern, and *time window* which is the maximum number of days that can occur between events to quality as an event sequence. For this illustrative analysis, we chose a support of 500 (at least 500 patients must have each pattern for it to be recorded) and an event window of 30 days (events must occur within 30 days in order to be considered part of the same sequence). With these parameters, running sequence mining on this dataset takes approximately 40 minutes[2] and yields 452 frequent sequences.

If we believe the results of the sequence mining algorithms contain the most salient features, then the problem goes from visualizing 900,000 events spread across 35,000 timelines into only 452 frequent sequences. We recently have built a prototype system, *Peekquence* [2], to begin exploring some of these open issues. While this work is still in its initial phase, it provides a playground to explore some of these concepts.

The core visual unit of Peekquence are mined patterns, rather than events. As patterns may contain many different event types and be composed of long event sequences, visualization techniques based on sankey diagrams (a la CareFlow [4]) or aggregated vertical bars

- *Adam Perer is with IBM T.J. Watson Research Center. E-mail: adam.perer@us.ibm.com.*
- *Bum Chul Kwon is with IBM T.J. Watson Research Center. E-mail: bumchul.kwon@us.ibm.com.*
- *Janu Verma is with IBM T.J. Watson Research Center. E-mail: jverma@us.ibm.com.*

Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis. Available online at: http://eventevent.github.io

[1] International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Available at www.cdc.gov/nchs/icd/icd9cm.htm.ICD-9

[2] Computed with a 2.5 GHz Intel Core i7 Processor and 16 GB of RAM

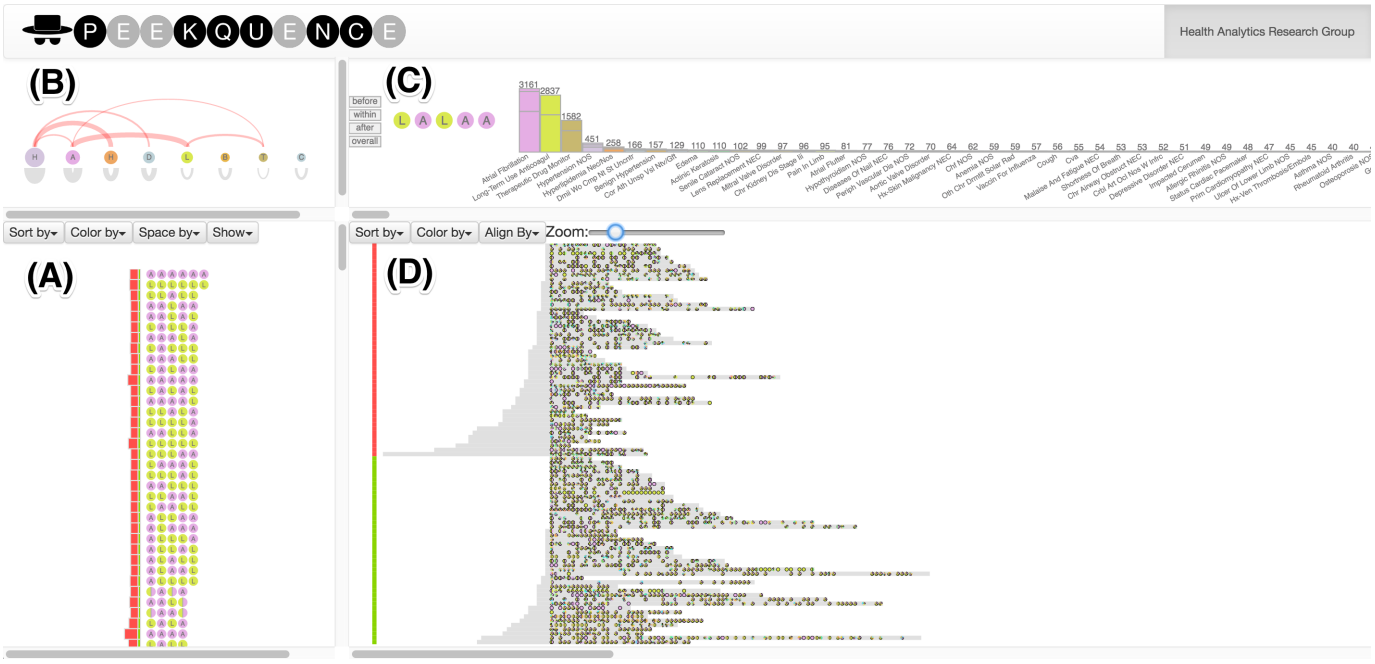

Fig. 1: Peekquence consists of four views: (A) the pattern list view showing patterns mined from SPAM with event sequences (colored circles with letters) as well as bars of patients with the ratio of case and control labels (diagnosis of a disease); (B) the sequence network view showing the frequency of event type co-occurrences within mined patterns; (C) the event co-occurrence histogram view showing the frequency of events co-occurring for a selected pattern; (D) the patient timeline view showing patients' event sequences that match the selected pattern.

(a la EventFlow [3]) tend to suffer from visual complexity without user-controlled filters based on domain expertise. Instead, we opted for a simpler visualization technique: a list of patterns, made up of *event glyphs* that visually encode each event type in the pattern. The event glyphs are visually encoded as circles, colored according to an categorical ontology, and labeled with an abbreviation of the event type's name. All of the four views in Peekquence, shown in Figure 1, use this glyph as the common visual element. In addition to a list of patterns (Figure 1A), there is an overview of common event types in the patterns (Figure 1B), histograms that summarize event types that co-occur with the patterns (Figure 1C), and a coordinated view to the actual patient timelines to understand how the mined patterns manifest in the actual data (Figure 1D).

Peekquence has led to interesting discoveries of the benefits and problems with relying on mined patterns as the main unit of visualization. There was no data curation done to the event types loaded into the user interface, but the algorithm was able to surface highly relevant types due to their prominence among patients with CHF (e.g. Atrial Fibrillation, Hypertension, and Hyperlipidemia – all common co-morbidities for patients with heart failure).

However, there are still many open issues in relying purely on automated techniques that may be problematic. For instance, the results of the mining algorithm treats patterns that have very subtle differences as completely different patterns (A →B →C) vs (A →B →B →C), and these subtle differences can account for a large percentage of patterns. Clustering techniques may prove useful to group together very similar patterns, but summarization of such clusters can be an additional challenge, as well as unintentional obfuscation of important patterns. Also, the patterns mined may include many irrelevant types of events that may add noise and reduce the signal of interesting patterns. Users could also distrust the mining results and lose confidence with the visual analytics system if they discover lots of uninteresting or irrelevant mined patterns. Finally, if users do not properly configure the parameters of the mining algorithms, results can vary. For instance, patterns that are not frequent as the user-specified support threshold will always be ignored. We need tools to illustrate the impact of each of these parameters so that their impact is transparent to users. Furthermore, there is a challenge of visually showing the occurrence of mined patterns within patients' records. It is necessary to integrate effective interaction techniques with visual analytics systems to support users' investigation from patterns (overview) to patients' records (detail).

Although we believe there are clear benefits to including analysis from data mining techniques, we are not advocating a purely automated approach. Instead, interactive systems for exploring temporal data should offer mining capabilities when the visualization technique cannot scale to the properties of the data. We propose developing guidelines to understand when data mining techniques are critical for analysis.

References

[1] J. Ayres, J. E. Gehrke, T. Yiu, and J. Flannick. Sequential pattern mining using bitmaps. pp. 429–435, 2002.

[2] B. C. Kwon, J. Verma, and A. Perer. Peekquence: Visual analytics for event sequence data. *ACM SIGKDD Workshop on Interactive Data Exploration and Analytics (IDEA 2016)*, 2016.

[3] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal Event Sequence Simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19, 2013.

[4] A. Perer and D. Gotz. Data-driven Exploration of Care Plans for Patients. In *ACM Human Factors in Computing Systems*, pp. 439–444, 2013. doi: 10.1145/2468356.2468434

[5] A. Perer and F. Wang. Frequence: Interactive mining and visualization of temporal frequent event sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pp. 153–162. ACM, New York, NY, USA, 2014.

[6] A. Perer, F. Wang, and J. Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of Biomedical Informatics*, 56(C):369–378, Aug. 2015.

[7] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. LifeLines: using visualization to enhance navigation and analysis of patient records. *Proceedings of the AMIA Symposium*, pp. 76–80, 1998.

[8] P. J. Polack Jr, S.-T. Chen, M. Kahng, M. Sharmin, and D. H. Chau. Timestitch: Interactive multi-focus cohort discovery and comparison. In *IEEE Proceedings of the Visual Analytics Science and Technology (VAST)*, pp. 209–210. IEEE, 2015.