# Exploratory Search Over Temporal Event Sequences: Novel Requirements, Operations, and a Process Model

**Taowei David Wang, Krist Wongsuphasawat, Catherine Plaisant, and Ben Shneiderman**
Department of Computer Science
University of Maryland, College Park, MD 20742
{tw7, kristw, plaisant, ben}@cs.umd.edu

**ABSTRACT**
Developing a detailed requirement analysis facilitates the building of interactive visualization systems that support exploratory analysis of multiple temporal event sequences. We discuss our experiences with collaborators in several domains on how they have used our systems and present a process model for exploratory search as the generalization of our experiences. This process model is intended as an outline of high-level analysis activities, and we hope can be a useful model for future and on-going exploratory search tools.

**INTRODUCTION**
Developing hypotheses about relationships among temporal events and assessing their plausibility are important exploratory tasks in a variety of domains. These tasks can be broken down roughly in two parts: (1) discovering notable event sequences, and (2) evaluating the prevalence of such sequences to strengthen analysts' confidence in their hypotheses.

To this end, several interactive visualization approaches have been proposed to support exploratory analysis in temporal event sequences: business intelligence and financial fraud detection [6], clinical care and medical research [1][3][4][10], and web session logs [2]. These approaches seek to solve the problems analysts face when using a command-line query interface or a pure data-mining approach. However, these approaches have significant differences in their support for interactive exploratory analysis. In particular, they have different support for aggregation, comparison, and advanced exploratory search features over temporal categorical data.

This paper focuses on analysis tasks, requirements, and designs for event sequences (e.g. database of electronic health records that contain diagnoses, treatments, interventions, and admission/discharge information, etc.) We introduce two prototype visualization systems: Lifelines2 [9][10] (Figure 1) and Similan [11] (Figure 2). Because the two systems are at different stages of development, and apply different strategies, they support different requirements. We discuss the requirements for exploratory analysis over this type of data, and how these systems address these requirements. We then discuss how our case study users utilize these strategies. Finally, we draw from our users' experiences to present a preliminary process model of information seeking in the context of event histories.

**SENTINEL EVENTS, ALIGN, RANK, AND FILTER**
In many situations, domain analysts have a question regarding a particular event. We call this central event "sentinel event". Analysts may seek (1) what are the most commonly occurring events immediately prior to or after the sentinel event, (2) what is the distribution of another event with respect to the sentinel event, (3) or study the length of time between a sentinel event and another event. For example, clinical researchers may be interested in the distribution of mammogram procedures in all patients, prior to their diagnosis of breast cancer, and also seek the average length of time between first diagnosis of cancer and the time of death is.

However, visualizations typically do not provide analysts a way to rearrange the data around sentinel events for a more effective presentation. Instead, the data is often fixed on a linear time line, making sentinel events, which can occur anywhere, hard to spot.

To address this problem, we designed the *alignment* operator. Alignment allows analysts to dynamically re-center the data around a sentinel event across all event histories. This allows patterns specific to the sentinel event stand out. In Figure 1, all histories are centered on the sentinel event *1st Radiology Contrast* (yellow triangles), obviates all events around the sentinel event. When histories are aligned, the calendar is set to be relative to the alignment instead of on absolute dates.

In Lifelines2 and Similan, analysts can specify a sentinel event by choosing the $n^{th}$ first or last event of a certain type. Additionally, they can also specify all events of a certain type to be all be sentinel events. This multiple alignment allows analysts to study distribution of events near to all occurrences of a specific type.

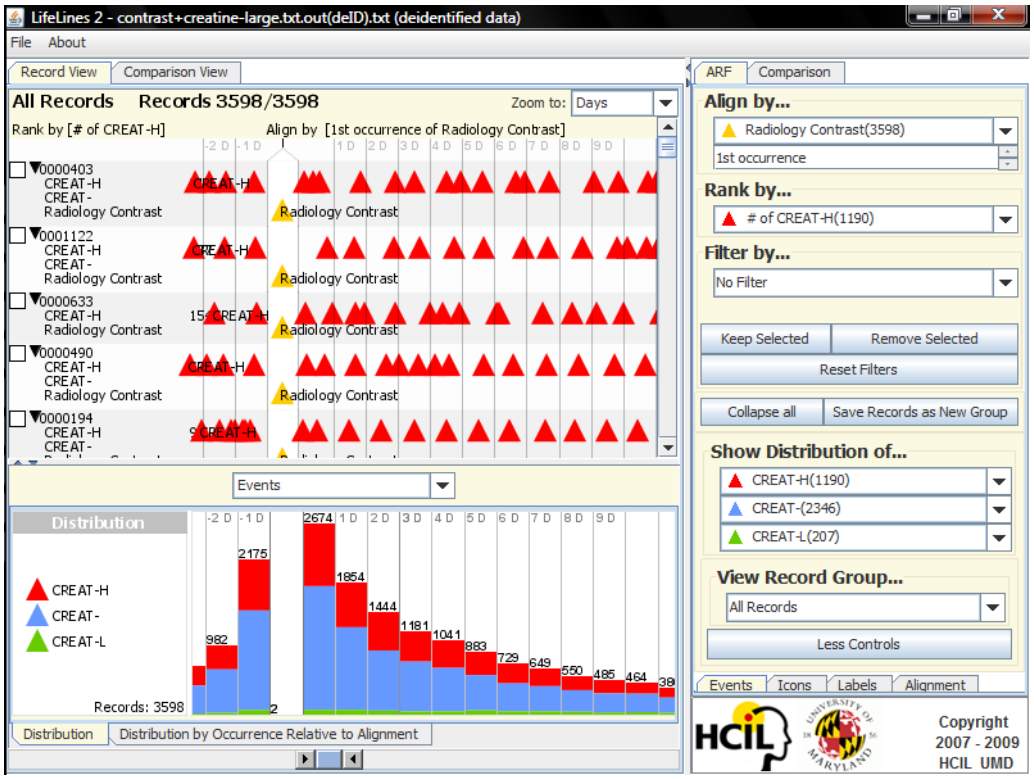


**Figure 1.** Screen shot of the Lifelines2 interface. The right portion is the control panel for a variety of operators. Top left is the main visualization panel, where each event history is shown as a horizontal strip on a time line. Each individual event is shown as a color-coded triangle (one event type is one color). The view shows that all histories are aligned by "Radiology Contrast" (the yellow triangles). The bottom half shows the temporal summary view of the red, blue, and green events over the visible time frame.

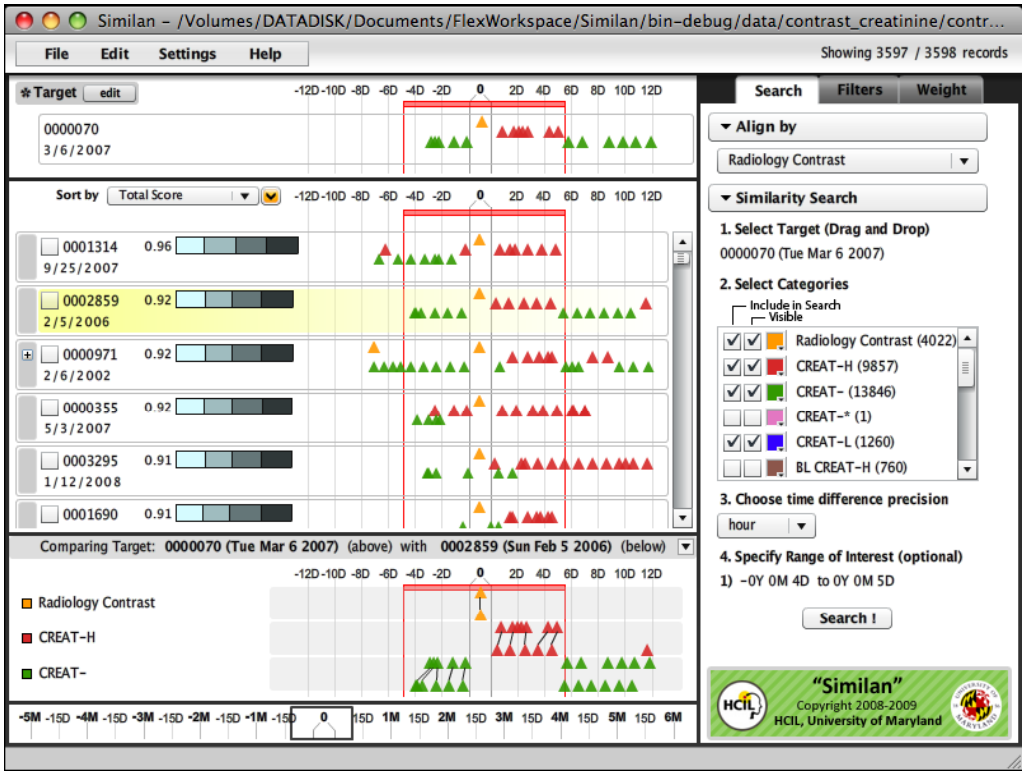


**Figure 2.** Screen shot of Similan. The right portion is the control panel. The left portion contains three major panels. The center panel is the visualization of all event histories. The top panel shows the target history the user has selected. All histories in the center panel are ranked by their similarity to the target. The similarity scores are represented by color-coded bars. The bottom panel shows the comparison between the target against a currently selected history (shown in yellow background in the center panel). The user has selected a timeframe (red rectangular region) over which the match algorithm operates.

In Lifelines2, the alignment operator is complemented with more traditional information visualization operators: rank and filter. Analysts can rank all event histories by, for example, the number of occurrences of high-blood pressure diagnoses, reordering the most severe patients to be on the top of the list.

There are two modes for filter. Analysts can filter in the similar manner as rank by specifying a number of occurrences of a specified event type. All histories that do not have at least that number of that event type will be filtered out. Analysts can optionally designate the occurrences of these events to be only before or after a

sentinel event. Secondly, analysts can specify a pattern of events to filter out histories that do not contain such pattern in an efficient manner [8]. An event pattern is a temporally ordered sequence of events or absence of events that analysts are interested. For example, analysts can use filter to find all patients who "were diagnosed with high-blood pressure, followed by no diagnosis of heart attack before a stroke."

## FINDING SIMILAR TEMPORAL EVENT SEQUENCES

The align, rank, and filter are the basic operators that allow analysts to study events of high interest and to find related events. However, sometimes analysts are interested in finding temporal event sequences that are similar to a specific history. For example, when a physician encounters a patient with symptoms that are rare and treatment options unknown, the physician may want to find past patients who share similar symptoms or medical history, and investigate the outcomes of different treatments.

This specific type of search has two main components. Analysts must specify what portion of a history is important, and what similarity means. In Similan, analysts would first picks a target history, and then choose a range on the time line to select a portion of that history that is relevant. The similarity matching is broken down to two parts. Similan first uses the Hungarian Algorithm [11] see how each history best matches the target. After the matches are found, Similan then assigns a similarity measure based on the number of mismatches and the "cost" of the match (based on temporal distance). Analysts can adjust the importance of mismatches. Analysts can also adjust the importance of out-of-order matches or matches with a large temporal differential.

Every history is then assigned a similarity measure, and displayed in descending order so that the most similar ones are on the top of the list. This is similar in spirit to the Rank-by-Feature framework, and allows analysts to review all histories before fine-tuning their search criteria. Analysts can review a similarity search and adjust the parameters of the similarity measure as described above to better suit their purposes.

The similarity search is further augmented to support "custom records". This means that analysts can manually specify a pattern to search instead of having to find one from an existing history.

## GROUPING, SUMMARY, AND COMPARISON

A natural extension to the variety of search mechanisms is to form subsets of histories for comparison. For example, hospital administrators may compare the differences of red blood cell counts for emergency room patients who experienced trauma and those who had not.

In Lifelines2, result of any filter operation can be explicitly made into a group. Analysts can choose to view any existing groups. They can also aggregate events for each group by using temporal summaries. Temporal summaries are stacked bar charts, where each stack represents one event type, aggregated over all histories. Analysts can examine the distribution of multiple event types at a glance [9]. The summaries are naturally integrated with alignment, so analysts can examine aggregations with respect to sentinel events.

Using temporal summaries, analysts can perform comparison among groups. A typical usage is to create two mutually exclusive groups and then put them side-by-side to study the temporal trend differences. A second use case is to successively narrow down a group of event histories and create successively smaller groups. Examining these groups' summaries gives analysts insight on whether this exploratory search path is on the right track.

## THE EXPLORATORY PROCESS MODEL

From working with our collaborators in medicine, student academic records, and law enforcement on drug trafficking phone records, we offer a preliminary process model of how our collaborators use our information visualization systems. Although the preliminary process model has numbered steps, our collaborators typically traverse in steps 2-4 in a pattern that is often not sequential.

1. Examine data for confidence (overview)
2. Exploratory Search
    a. Iteratively apply visual operators
    b. Evaluate results of manipulation
    c. Deal with unexpected discoveries
3. Analysis, Explanation
    a. Examine paths of search as a whole
    b. Determine to what extend are the questions answered
        i. At the limitation of the system
        ii. At the limitation of the data
    c. Refine existing questions
4. Report results to colleagues
    a. Document findings
    b. Disseminate subsets of data
5. Move onto new questions

One of the most common results of users looking at their own data through a visualization technique for the first time is the surprise that there are artifacts in the data (systematic errors, lack of consistency, etc.). This is because they have never seen it in an effective format before. As such, our collaborators would cursorily browse the data to make sure the data reflects what they know.

After gaining confidence of the visualization and of the data, they would start seeking answers to their pre-conceived questions. However, new questions often

spawn when they notice interesting or unexpected data. At this point they would utilize their domain knowledge to try to explain what they see, or they would write down the new question for later exploration. We noticed that analysts may apply alignment on different sentinel events in the same exploratory session to look at the data in different views. They would actively manipulate the display by ranking, filtering iteratively, or change how similarity is weighted in Similan's search. However, alignment remains the strongest indicator on what focus they have on the data.

We found that aggregation techniques such as temporal summaries allow the analysts to look at the data quickly. Many of them learned to visually focus only on the summaries. They would also inspect the previously created groups by comparing their summaries to see qualitatively what kind of progress they have been making, and decide whether the path they are taking has potential. When they see a view of the data that answers their questions or contain interesting discoveries, they would save the state of their progress – saving the current group, and taking screen shots.

Although the process model we present here is still very preliminary, it already suggests elements that are indispensible to exploratory search in temporal categorical data. The first is a way to "anchor" the visualization for a particular path of search (like *alignment*), and allow analysts to quickly and dynamically change the anchor. The second is an overview of the entire dataset so that a mental model can be built quickly as the data is being manipulated. Next, a way to explicitly track users' steps of exploration is important. Finally, features that support viewing and comparison of different steps of exploration are critical to backtracking and taking excursions in the search process. We recommend these features for future applications.

## DISCUSSION

Performing exploratory analyses using a command-line query tool suffer from the problem that users have no mental model of the data. As a result, users have a hard time making judgments on how to refine their exploratory steps. Similarly, in a pure data-mining approach, lack of a mental model of the data makes interpretation of the results tricky. Information visualization allows opportunities for users to orient themselves at each step of the exploratory search, and enables maintenance of a consistent mental model throughout the process.

This paper presents several visualization and interaction techniques to let users control their exploratory paths and sustain a working mental model in searching temporal events. We argue that these approaches are more amenable to exploratory search. Information retrieval applications on temporal data can leverage work presented here to provide users a more fulfilling search experience. We discuss a preliminary process model for event sequences, and we hope to see interactive visualization techniques to be used in conjunction with information retrieval or data mining techniques to connect to their users as in [5][6].


## ACKNOWLEGEMENT

This project is supported in part by the Washington Hospital Center and Harvard - Partners HealthCare.


## REFERENCES


1. Fails, J. Karlson, A., Shahamat, L., and Shneiderman, B., A visual interface for multivariate temporal data: finding patterns of events over time", *Proc.IEEE VAST*, 2006

2. Lam, H., Russell, D. M., Tang, D.,Munzner, T., Session viewer: supporting visual exploratory analysis of web session logs. *Proc. IEEE VAST*, 2007

3. Plaisant, C., Lam, S., Shneiderman, B., Smith, M., Roseman, D., Marchand, G., Gillam, M., Feied, C., Handler, J., and Rappaport, H., Searching electronic health records for temporal patterns in patient histories: a case study with microsoft amalga," *Proc. AMIA Annual Fall Symposium*, 2008.

4. Ong, J., DataMontage Software, http://www.stottlerhenke.com/datamontage, 2006.

5. Post, A. R., Harrison, J. H.. Protempa: A method for specifying and identifying temporal sequences in retrospective data for patient selection. *JAMIA*, 2007.

6. Shahar, Y., Cheng, C., Intelligent visualization and exploration of time-oriented clinical data. *Proc. HICSS 1999*.

7. Suntinger, M., Schiefer J., Obweger H., and Groller, M.E. The event tunnel: interactive visualization of complex event streams for business process pattern analysis. *Pros. of IEEE Pacific Visualization Symposium '08*, 111-118, 2008.

8. Wang, T.D., Deshpande, A., Shneiderman, B., A temporal pattern search algorithm for personal histories, Tech Report # HCIL-2009-14, 2009, http://hcil.cs.umd.edu/trs/2009-14/2009-14.pdf.

9. Wang, T.D., Plaisant, C, Shneiderman, B., Spring, N., Roseman, D., Marchand G., Mukherjee, V., and Smith, M., Temporal summaries: supporting temporal categorical searching, aggregation and comparison. To appear in *Proc. IEEE Infovis, 2009*.

10. Wang, T.D., Plaisant C, Quinn, A., Stanchak, R., Shneiderman B., and Murphy, S., Aligning temporal data by sentinel events: discovering patterns in electronic health records. *Proc. CHI, 2008*.

11. Wongsuphasawat K. and Shneiderman B., finding comparable temporal categorical records: a similarity measure with an interactive visualization. To appear in *Proc. IEEE VAST, 2009*.