Interactive Poster: Visual Data Mining of Unevenly-Spaced Event Sequences

Alex Godwin* Remco Chang† Robert Kosara‡ William Ribarsky§

Visualization Center
University of North Carolina at Charlotte

Abstract

We present a process for the exploration and analysis of large databases of events. A typical database is characterized by the sequential actions of a number of individual entities. These entities can be compared by their similarities in sequence and changes in sequence over time. The correlation of two sequences can provide important clues as to the possibility of a connection between the responsible entities, but an analyst might not be able to specify the type of connection sought prior to examination. Our process incorporates extensive automated calculation and data mining but permits diversity of analysis by providing visualization of results at multiple levels, taking advantage of human intuition and visual processing to generate avenues of inquiry.

1 Introduction

An alignment between two sequences is defined as the elements that are found in both sequences in the same order, potentially with gaps between the elements. For example, if two patrons of a movie rental chain each rent the same three movies in the same order, those three movies, or rental events, represent an alignment between the sequences of rental events for the two customers. This is true even if there were significant gaps between matching rentals or if many of the movies each customer rented do not match. Sequences alignments can be computed by the longest common subsequence (LCS) algorithm, and once an alignment is determined it can be scored based on the number of gaps or unmatched events. Traditionally the LCS algorithm does not penalize the score of an alignment if there are temporal gaps between aligned events, and it does not incorporate the difference in time between matching events.

Our method uses an LCS algorithm altered specifically to incorporate time as a parameter for sequence comparison. The modified LCS allows the user to specify the penalties imposed on an alignment score for gaps in time between matched events or penalties for matched elements that occurred at significantly different time periods. One of the interesting properties of the LCS solution is that the alignment between two sequences can be scored without expending the additional computation required to provide a detailed representation of it. This property allows us to mine the data first for implicit relationships, tune the parameters, and drill into the details of each alignment once attention is focused. In this manner a user can query thousands of entities for potentially interesting results before exerting the extra time and effort to dissect discovered alignments. Mining the data with visual interaction facilitates the incorporation of human feedback into the exploration process, particularly when little is known about the data and the exploration goals are vague [4].

2 Overview of All Database Alignments

The user begins the process of analysis by running a pairwise LCS of all of the entities in the database, creating a two dimensional score table. The scoring table is then used to generate two overviews, each supporting interactions to explore the structure of the database.

2.1 Pixel-Oriented Overview

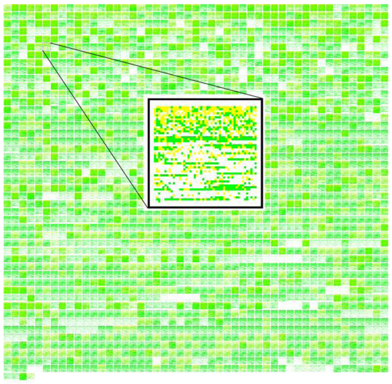


Figure 1: Pixel-oriented overview of thousands of sequence alignment scores.

Dense, pixel-oriented displays are an efficient representation of data that allows a massive number of values to be portrayed on screen simultaneously [3]. The pixel-oriented overview consists of thumbnails for each of the entities in the database laid out in a two-level hierarchy (figure 1). Each pixel within a thumbnail is colored to represent the magnitude of the alignment score between the entity that the thumbnail represents and the entity in the corresponding position in the top-level view.

Sorting and brushing operations, combined with filter and highlight, reveal associated entities and allow the user to explore the associations between groups of entities. A slider widget allows the user to set the color gradient for scores using a graphical representation of the distribution of the scores. Additionally, a user may highlight a thumbnail and all of the thumbnails that contain the same pattern of pixel coloration. The highlighted thumbnails then represent a connected structure of potentially related groups. This is one way in which networks of similar entities can be revealed in the database.

2.2 Social Network View

A second overview is provided that directly maps the scores between entities as a force-directed graph, created in the prefuse toolkit. In this representation, entities (represented by floating nodes) are connected by an edge when their alignment score is

*e-mail: jagodwin@uncc.edu
†e-mail: rchang@uncc.edu
‡e-mail: rkosara@uncc.edu
§e-mail: ribarsky@uncc.edu

above a user-defined threshold. Nodes that are unconnected by edges have a repel force, while connected nodes attract one another. This confluence of forces causes the graph to untangle over time, revealing any present structure (figure 2).

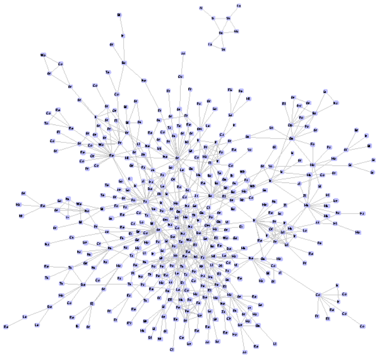
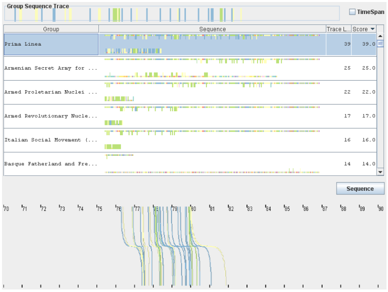


Figure 2: The sequence comparison scores have been used to generate a force-directed graph. The visualization indicates structured subnetworks of aligned groups as well as satellite subnetworks.

3 Detailed Alignment View

Once the user has identified an entity to explore further it can be selected in the detail viewer. The chosen entity is compared to every other in the database, and the results are then displayed in a table that can be sorted. Each row of the table contains an event index representation of the alignment between the selected entity and one other entity in the database. Selecting a row reveals the time spline view for the indicated alignment.



Group	Sequence	Trace L...	Score ▾
Prima Linea		39	39.0
Armenian Secret Army for ...		25	25.0
Armed Proletarian Nuclei ...		22	22.0
Armed Revolutionary Nucle...		17	17.0
Italian Social Movement (...		16	16.0
Basque Fatherland and Fre...		14	14.0

Figure 3: Detailed comparison views of one sequence queried against the database.

Event Index View

The event index view reveals the alignment between the selected sequences and each of the other entities in the database(figure 4). This view preserves the order and cardinality of the sequences but not the length of time between events [1]. The sequence of events for the selected entity is displayed at the top of the image, and the sequence for the entity against which it's being compared is along the bottom. If the event is an aligned element, it is promoted in size. The color of each event represents its value, so if all the aligned events are the same, as in two shoppers who only buy the same object every trip, then it will be apparent. Using color, diverse alignments and shifts in type of aligned events can be recognized.

Time-Spline View

The time-spline view (figure 3) gives a representation of the alignment that preserves the position of each aligned event in time. The time splines connect the position of each aligned event in the selected group's sequence (along the top of the image) to the corresponding position in the time line of the secondary group.

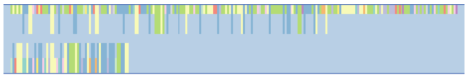


Figure 4: Closeup of the event index detailed view.

4 Data Studies

Global Terrorism Database

We began to explore the usefulness of this tool using the Global Terrorism Database(GTD), an open source database from the University of Maryland's Center for the Study Terrorism and Responses to Terrorism (START). The GTD consists of more than sixty thousand worldwide terrorism events recorded from 1970 to 1997, and details the actions of more than 2,300 terrorist groups. The use of our technique with the GTD has provided encouraging results, as it has provided insight into seemingly unrelated groups that, once researched, had verifiable connections [2]. We hope to further explore this database with the aid of expert evaluation.

Wire Transactions

To illustrate the diversity of our methods for analysis, we've begun to test its effectiveness on databases consisting of wire transactions of funds between banks. The thousands of transfers in the database span a period of one year and contain information for the sender, receiver, and any keywords associated with the transaction. Our technique quickly highlights potential trends in keywords and connections between entities, but due to the highly sanitized nature of the data it's difficult to confidently state the level of certainty in these comparisons without further investigation. However, the potential gain in yielding significant results in this application area represents an exciting possibility.

5 Conclusions and Future Work

We have presented a novel analysis method for combining algorithmic data processing with multiple views to support the intuitive transition from hypothesis generation to verification of results. This method is particularly suited for the chosen alignment algorithm, a modified LCS. To solidify the significance of this contribution, further study will be necessary to determine the ease of user interaction, as well as enhancing the results of a query with reported statistical significance.

Acknowledgements

This work was performed with partial support from the National Visualization and Analytics Center (NVAC), a U.S. Department of Homeland Security Program, under the auspices of the SouthEast Regional Visualization and Analytics Center. NVAC is operated by the Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy Office of Science laboratory.

References

[1] A. Aris, B. Shneiderman, C. Plaisant, G. Shmueli, and W. Jank. Representing unevenly-spaced time series data for visualization and interactive exploration. In *Proceedings of the International Conference on Human-Computer Interaction (INTERACT)*, pages 835–846, 2005.

[2] A. Godwin, R. Chang, R. Kosara, and W. Ribarsky. Visual analysis of entity relationships in global terrorism database. In *SPIE Defense and Security Symposium*, 2008.

[3] D. A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

[4] D. A. Keim. Information visualization and visual data mining. *Transactions on Visualization and Computer Graphics*, 7(1):100–107, 2002.