Geovisualization of Spatio-Temporal Events in STempo

Brian Swedberg, Anthony Robinson, Frank Hardisty, Donna Peuquet

Abstract — It remains difficult to develop a clear understanding of geo-located events and their relationships to one another, particularly true when it comes to identifying patterns of events in less-structured textual sources, such as news feeds and social media streams. Here we present a design for a geovisualization tool that can leverage computational methods, such as T-pattern analysis, for extracting and analysing patterns of interest from event data streams. Our system, called STempo, includes a coordinated-view geovisualization designed to support visual exploration and analysis of event patterns in terms of time, geography, and content. Through a study example, we highlight the utility of STempo for understanding patterns of political, social, economic, and military events in Yemen during the 2011 Arab Spring uprising.

Index Terms — event data; geovisualization; T-pattern, temporal analysis, sequence analysis

Introduction

The unforeseen and unprecedented political, civil, and military upheavals in the Middle East during 2010 and onward, informally known as the Arab Spring, has highlighted the complexity involved in making sense of event data. Analysts can no longer rely on traditionally held cause and effect beliefs, as new patterns form and old patterns may evolve. Although spatially grounded event data sources are more readily available than ever before, it remains a challenge to explore and reason about sequences of events. This fast paced climate prompts a need for methods that will aid analysts in identifying and recognizing space-time patterns in event data. Specifically, new computationally-enabled approaches are needed to help identify which sequences of events, apart from thousands of possibilities, may be of most interest.

To support visual exploration and analysis of complex patterns of events derived from large spatio-temporal event datasets, we have developed STempo, a coordinated-view geovisualization and geocomputation application (see main panel in Figure 1). STempo leverages TABARI and CAMEO systems to automate event categorization from natural language text input [1] and T-pattern analysis to extract event patterns [2]. But the standard architecture of TABARI/CAMEO is designed to support a collection of different categorization, and T-pattern analysis is intended for use with data from controlled, laboratory settings. We therefore made significant changes to both. Our objective is to provide analysts with a tool to identify event patterns, generate new hypotheses, and support the confirmation or denial of previously held hypotheses using real-world event data. Through a case study of Yemen during the Arab Spring we have found STempo to be an effective solution. We plan to continue refinement of geovisualization capabilities within STempo via a user study, and work toward scaling-up for handling extremely large collections of event data from Really Simple Syndication (RSS) news feeds.

1 Related Work

Our goal is to support exploration and analysis of large space-time event datasets without the need for maintaining a clear hypothesis about which canonical sequences may be important to evaluate. We build here on a multitude of previous efforts to design and develop coordinated-view geovisualizations to explore and analyze spatio-temporal datasets. The computational foundations for our current work grow from significant advances in pattern recognition techniques in the field of data mining, and specifically from knowledge discovery from database (KDD) techniques demonstrated in the context of GIScience.

Some of the most promising methods for investigating event data

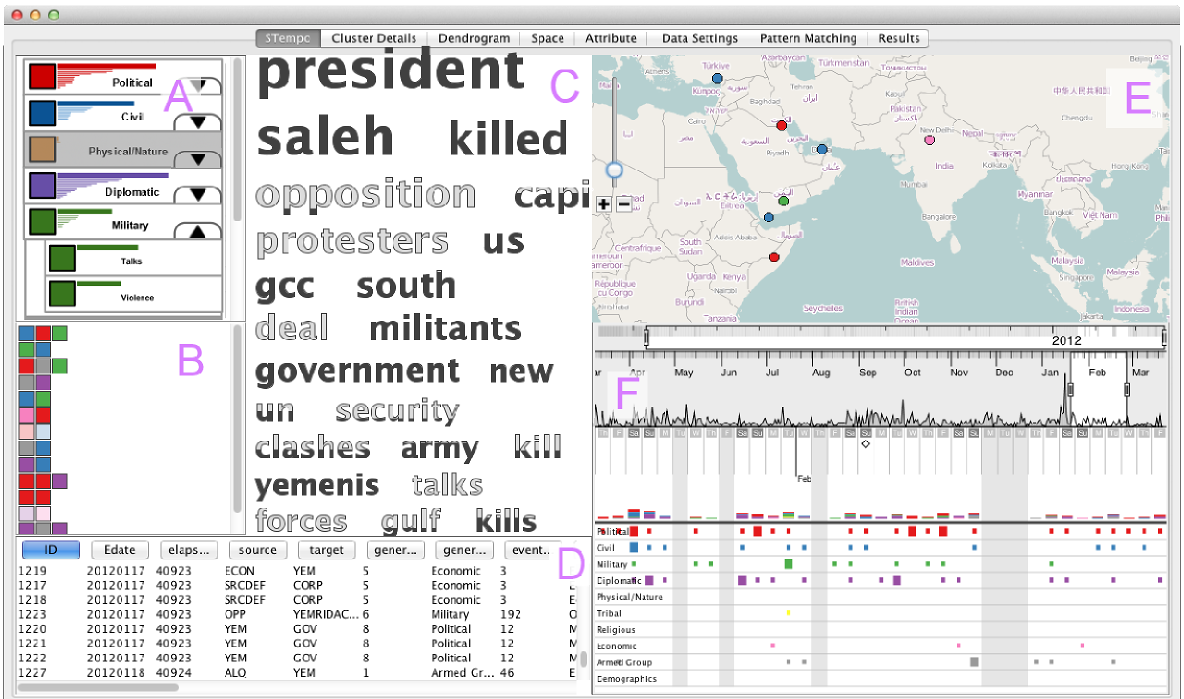

Fig. 1. STempo main panel

stem from a combined geovisualization/computation approach. For example, [3] uses genetic sequencing methods to explore user interactions with geovisualizations, and [4] performs pattern analysis from spatial movement data within a visualization environment. Our own research builds on a previous project to analyze spatio-temporal data and to uncover meaningful sequences in space-time events using geovisualization [5]. The work reported upon here continues from these examples, with an approach for discovering and visualizing spatial event sequences.

2 STempo

To support the discovery and visual exploration of pattern results from space-time event data, we developed the STempo geovisualization environment. STempo provides a wide range of coordinated-view displays with which users can investigate space, time, and attribute data for to uncover patterns and phenomena. STempo is built using a polyglot programming strategy that integrates capabilities from the Java and Python languages. In the following sections we describe our event data categorization, pattern discovery, and interface design.

2.1 Event Data

RSS feeds and news articles contain valuable information that could help answer questions like, ‘Which are the new, as well as which are the unexpected, patterns that emerge from the *Arab Spring* events?’ and ‘To what extent do social, economic, political, and other types of events occurring at a location exhibit consistent patterns?’ As text analysis is a complicated subject and not the emphasis of our research, we require pre-processing prior to importing data into our geovisualization environment to classify and to approximate the geographic locations of individual news reports.

For or analyses we first removed any obvious redundancies in news reports relating to a given event, and encoded event types for each non-redundant report based on the title of each news event using the TABARI and CAMEO categorization tools. TABARI is an open-source software package for identifying and categorizing events, and CAMEO provides associated dictionaries that are specifically tailored for political, diplomatic, and military applications [1]. TABARI and CAMEO were built for the purpose of detecting insurgencies through political and civil events, so for our goal of discovering broader spatio-temporal patterns, we extended the dictionaries to include such categories as economic, religious, and entertainment/sporting events. We also found that some built-in CAMEO event types (i.e., ‘reject’ and ‘disapprove.’) are overly vague for the purposes of our case study, since, for example, a foreign government voicing reaction to military actions in Yemen is very different in nature from reports of civilians in Yemen reacting to win/loss of a soccer game. We therefore implemented a post-processing program to further specify event types. Focusing on the actors for each event as identified by TABARI, each event is encoded by one of ten high-level event categories: *Armed Group*, *Civil*, *Demographic*, *Diplomatic*, *Economic*, *Military*, *Physical/Nature*, *Political*, *Religious*, and *Tribal*. With this added event type designation, we are able to distinguish between, for example, a political disapproval event – which occurs frequently – and diplomatic disapproval, which may have larger international consequences. Each article is encoded as a combination of high-level event type and detailed event type in the form *Political – Disapprove*, for event classification.

We gather the spatial content from our data in a two-step process. The HTML tags contained within the source code of web-based news reports are used to automatically identify and extract the body text. We then use part-of-speech tagging to determine geographic location names within the text.

2.2 Computational Component

Our approach leverages T-pattern analysis to identify event sequence patterns. T-Pattern analysis is a technique developed for application in psychological studies to expose previously unknown patterns of human behavior in the context of interpersonal social interactions. To identify significant co-occurrences of events from global event data sets, we have developed an extended version of the T-pattern analysis technique first described in [2]. The T-pattern process identifies time distances between multiple occurrences of two types of events, and groups event types that are not likely to occur by chance defined via a user-specified p-value. We have extended the original T-pattern method by implementing the means to further identify and ignore redundancies in the event data, ignore known spurious relationships, to incorporate locations as distinguishing characteristics of events (e.g. an episode of civil unrest in one city would be treated distinctly from an episode of civil unrest in another location), and to reveal the statistically significant relationships of all event types to a given event type. The end result of T-Pattern analysis a set of hierarchically-organized patterns of significantly related event types that re-occur in the same basic temporal sequence (i.e. *Tribal – Protests*, *Political – Issues Negative Statement*, *Military – Violence*).

We have built our system to allow plug-in capabilities for other pattern analysis processes. One such additional process currently implemented in STempo is pattern matching so that additional occurrences of ‘interesting’ patterns, once found by T-pattern analysis and selected by the user, can be identified in other places and/or times. Sequence alignment analysis originally developed to support genetic research has also been recently applied in GIScience to evaluate the effectiveness of geovisual analytics displays by evaluating sequences of recorded user interactions [6], and to compare trajectories of pedestrian tourist activity [7].

2.3 STempo Geovisualization Design

The STempo interface features multiple views designed to facilitate exploration and analysis of event data and event sequences. Tabs are used to organize multiple panels and each provides specific groupings of linked visualizations, as views. Throughout our interface we represent each event by its higher-level category using a qualitative color scheme from ColorBrewer [8]. Interaction is supported through the use of linked selections and highlighting [9] across views. These selections also persist when users switch tabs to view different view collections, thus making it possible to begin discovery using one set of views and continue interrogating those patterns in another set of views.

In our primary panel, shown in Figure 1, we display six views that include a spatial, temporal, and attribute representations of our event data.

2.3.1 Legend

Legends provide users with a graphical reference to data categories and attributes. Our system builds on this tradition by including a legend (Figure 1.a) that acts as reference as well as an interface to select and filter. In order to select or filter a category or sub-category, a user clicks on the associated legend item. We shade the background of each category based on the amount of selected sub-categories to reflect its respective selection status.

2.3.2 Sequence Overview

The ‘Sequence Overview’ (Figure 1.b) provides a simplified graphical summary of computational pattern analysis results. Sequences are displayed as a horizontal series of colored squares. Thus, each row represents a single pattern. This simplified representation of T-Pattern results reads left to right for any given pattern, and patterns are ranked, top to bottom, by decreasing statistical significance. For further information about a pattern, users may hover over the colored squares or investigate one of the secondary panels. A key affordance in this view is the ability for users to select a T-pattern and filter the rest of the display to show only the events related to that pattern.

2.3.3 Tag Cloud with SymWords

Our dynamic tag cloud (Figure 1.c) shows the key words derived from text used to categorize events. We introduce the SymWords concept and associated implementation in order to create a word cloud that is more analytically useful. SymWords (short for Symbolic Words) are sets of semantic letterforms that are systematically visually transformed in correspondence to data, essentially applying Bertin's visual variables to words, and particularly to the interior of the letterforms. The size of any particular SymWord corresponds to the frequency of occurrence for that word over all news stories within the entire data set, whereas the fill of each SymWord changes to represent the frequency of that word within the current selection. Furthermore, we have adopted an ordered view of these SymWords to allow for easier identification and comparisons, as opposed to popular methods like Wordle [10] which sacrifice such capabilities in exchange for aesthetically-pleasing renderings. Displaying the correspondence among the frequency of the terms in temporally selected sets of events is a powerful means of discovering and displaying trends over time. The ability to interactively eliminate stop words allows the analyst to focus on a particular pattern.

2.3.4 Table

Our table (Figure 1.d) displays the raw data records. Unlike other representations, the table has access to all attributes within the imported data set. In this view, users can sort and select elements stored within the original data table.

2.3.5 Map

Our interactive map (Figure 1.e) shows the location of each geocoded event, and events are colored to represent its associated higher-level category. In order to uncover information about a geolocated event, users can hover their mouse over any point to access event details, such as its date, brief description, and a hyperlink to the original source article.

2.3.6 Multipart Timeline

Following on earlier work that introduced the idea of hierarchical timelines [11], our timeline (Figure 1.f) provides users with a temporal representation of the data. The timeline is composed of three adjustable hierarchal views: the ‘Whole Global View’ (top), a ‘Coarse View’ (middle), and a ‘Fine View’ (bottom). The Whole Global View displays ticks along the x-axis to represent currently selected events within the data set. The Whole View also allows the user to adjust the content sent to Coarse and Fine View via handles. The Coarse View has similar display and functionality as the Whole Global View, however it displays a curve to show the overall frequency of events over time.

The Fine View aggregates all of the events by a period of time. In the case of our example data, this period of time is one day. Aggregates are colored according to the Legend: If some events are selected, then the background is colored white. If all events with are deselected then the background is colored light grey. If no events occur within the aggregate then the background is shaded dark grey.

We display three temporal representations of the aggregates within the Fine View: (a) a section showing notable events not in the data itself, such as civil or religious holidays, as hollow black diamonds, (b) stacked bar charts to display event categories by total frequency, and (c) graduated symbols to further represent event category frequency. In a manner similar to our Map, a popup appears whenever the mouse hovers over a graduated symbol. This popup provides a brief description and hyperlink for each event represented by the symbol.

2.3.7 Secondary Collections

Among several view collections accessible through tabbed panes in STempo, we designed a secondary pane for T-Pattern results. This pane contains a dendrogram and treemap—both coordinated to the primary pane. The dendrogram displays the hierarchical structure patterns resulting from the T-pattern analysis. Each horizontal branch indicates the expected amount of time between successive events using the length of color-coded branches. The color of each branch signifies the higher-level event type, as in the Sequence Overview. The treemap shows T-patterns as a function of the number of events associated with the unique elements. Much like many of the other views, popups are used to provide more details, such as, the p-value, event subcategory, and frequency.

3 Use Example

To test STempo’s capabilities, we captured RSS news feed data from 2011-2012 using a custom web-scraping software tool. We collected data from February 3, 2011 to March 21, 2012 and focused on events that either occurred in Yemen or in which an actor associated with Yemen played a role. After importing the data into STempo we immediately see, within the primary panel, the spike in event frequency around January 2012 within the Multipart Timeline. To investigate events that led to this phenomenon, we slowly drag the Timeline's 'Coarse View' selection window toward the spike to chronologically animate the linked views. While doing so we notice that the word 'protesters' appears to be in opposition of 'militants,' while the word 'killed' and 'president' are more sporadic. After arriving at the spike in frequency, we notice that the majority of these events are due to diplomatic events. As we continue to drag the selection panel down the timeline, we notice in the tag cloud the lack of fill for the word 'protesters' and total fill for the words 'militants', 'south', and 'army' from mid-January to the end of March (Figure 1), indicating event frequency in the selection. These findings raise several additional questions: Does the sudden increase in diplomatic events have anything to do with the rise and fall of protestor and militant related events? Do protestor events spur militant events, or are protesters suppressed by actions of militants? Which group is suffering the most casualties?

To further investigate, we first adjust the timeline handles to approximately one month before and one month after the spike. We then consult the ‘Sequence Overview’ to ground some of our suspicions. Among fourteen statistically significant patterns we are especially intrigued by the pattern: *Religious - Holiday/Holy-Day, Diplomatic - Arrest/Detain/Charge with Legal Action, Tribal - Abduct/Hijack, Political – Threaten*. By clicking on that pattern we drill-down to its associated events (Figure 2). Hovering over the timeline and map we open several news articles that relate to events identified as members of the T-Pattern.

We can piece together a hypothesis about the causes of the mid-January events with the help of pattern analysis, domain knowledge, and news articles. Our hypothesis involves the holy-day of Ashura, which occurred a month prior to the spike in observed events. Through domain knowledge we know that this holy-day causes tension between Sunni and Shia Muslims. Following Ashura we notice arrests and legal action events immediately before the spike. Perhaps heightened tension from the holy-day sparked more aggressive behavior? A short while later, we see several kidnappings. Through the use of the map and timeline we identify the articles that detail the kidnappings. One article states that the abductors are demanding the release of a person who is being held for minor crimes [12]. Based on this evidence we wonder if this systematic pattern influences kidnappings.

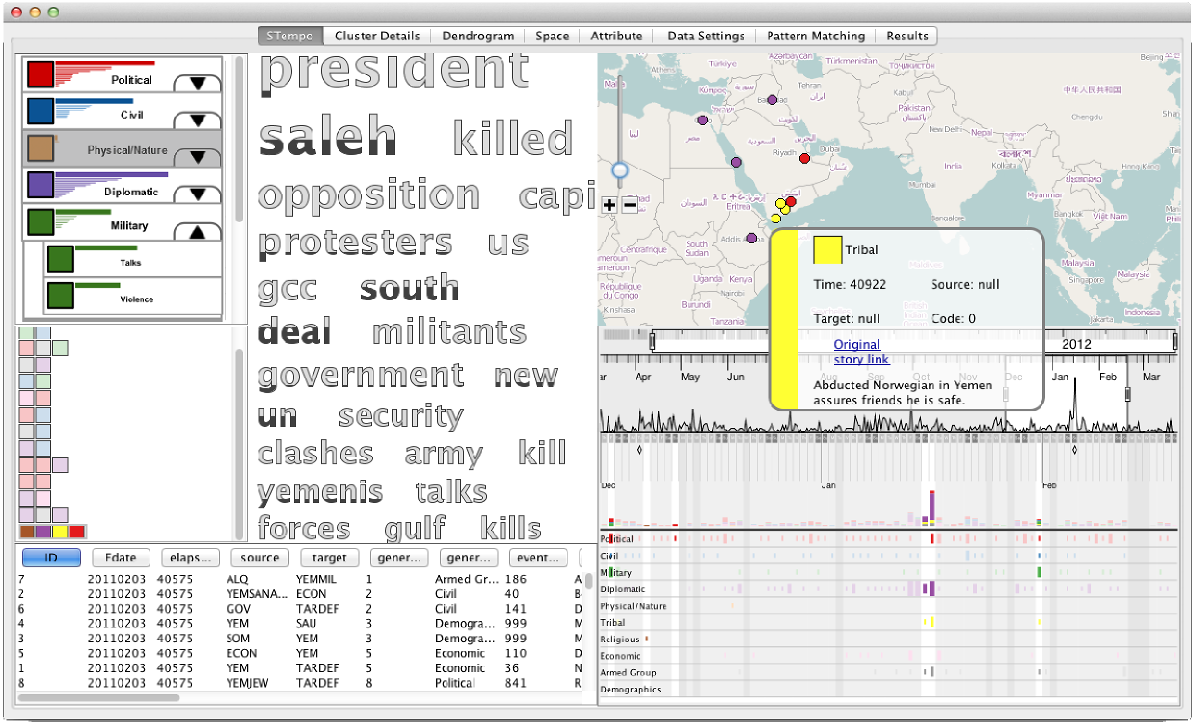

Fig. 2. STempo drill down analysis

4 Conclusion

In this paper we have presented STempo, a system for geovisualization and analysis of event data. We have outlined how computational methods can be paired with geovisualization to enhance spatio-temporal pattern analysis capabilities. We have also shown through a brief use example how STempo and our modified T-pattern analysis approach can be used to gain insight relating to patterns that may not have been previously predicted or readily understood.

We are planning some additional visualization refinements, including enhanced map symbolization to show details for multiple events that occur at a single location. A user evaluation of STempo is an essential next step in our further refinement of its tools and interactive affordances in support of analytical reasoning. While the toolkit has undergone multiple rounds of internal usability refinement and consultation with political science domain experts, we do not yet know if other analysts will understand T-pattern results in the way we present them, nor do we fully understand how such event sequences, once detected, would best make their way into an analyst's typical workflow. It will also be important to evaluate STempo with analysts to identify which combinations of views should be collected in each panel. We believe a potential strength of this system is its ability to assist analysts engaged in information foraging as part of the sensemaking process [13]. To better support this focus, we are currently improving view coordination through multiple modes of selection. Results from this effort will be a new collection of views to enable iterative evidence collection and annotation.

To further expand and explore the combined geovisualization/geocomputational possibilities for space-time event analysis in STempo, we are also in the process of greatly expanding the temporal and spatial scope of our data to include multiple decades and all countries of the Middle East. We are also greatly expanding our event type ontology to include a much broader and richer range of event types. News events continue to be captured going forward via our own web scraping software tool. Past news events are being captured via downloads of archived news feeds from Factiva.

References

[1] P. A. Schrodt, *et al.*, "The CAMEO (Conflict and Mediation Event Observations) actor coding framework," presented at the International Studies Association Annual Meeting, San Francisco, 2008.

[2] M. S. Magnusson, "Discovering hidden time patterns in behavior: T-patterns and their detection," *Behavior Research Methods, Instruments, & Computers,* vol. 32, pp. 93-110, 2000.

[3] S. Fabrikant, *et al.*, "Novel method to measure inference affordance in static small-multiple map displays representing dynamic processes," *The Cartographic Journal,* vol. 45, pp. 201-215, 2008.

[4] G. Andrienko, *et al.*, "From movement tracks through events to places: extracting and characterizing significant places from mobility data," presented at the IEEE Symposium on Visual Analytics Science and Technology (VAST 2011), Providence, RI, 2011.

[5] C. Weaver, *et al.*, "Visual exploration and analysis of historic hotel visits," *Information Visualization,* vol. 6, pp. 89-103, 2007.

[6] A. Çöltekin, *et al.*, "Exploring the efficiency of users' visual analytics strategies based on sequence analysis of eye movement recordings," *International Journal of Geographic Information Science,* vol. 24, pp. 1559-1575, 2010.

[7] N. Shoval and M. Isaacson, "Sequence Alignment as a method for human activity analysis in space and time," *Annals of the Association of American Geographers,* vol. 97, pp. 282-297, 2007.

[8] M. Harrower and C. A. Brewer, "Colorbrewer.org: an online tool for selecting color schemes for maps," *The Cartographic Journal,* vol. 40, pp. 27-37, 2003.

[9] A. C. Robinson, "Highlighting in geovisualization," *Cartographic and Geographic Information Sciences,* vol. 38, pp. 373-383, 2011.

[10] F. B. Viegas, *et al.*, "Participatory visualization with Wordle," *IEEE Transactions on Visualization and Computer Graphics,* vol. 15, pp. 1137-1144, 2009.

[11] D. F. Huynh. (2008, *SIMILE—Timeline." Simile Projects*. Available: http://www. simile-widgets.org/timeline/

[12] H. Almasmari. (2012, *6 aid workers kidnapped in Yemen." CNN*. Available: http://www.cnn.com/2012/01/31/world/meast/yemen-aid-workers/index.html/

[13] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *International Conference on Intelligence Analysis*, McLean, VA, 2005.