Validation of Results for Temporal Patterns Derived in STempo

D.J. Peuquet, S. Stehle

GeoVISTA Center, Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802, USA
Email: (Peuquet, Stehle)@psu.edu

1. Introduction

Deriving understanding from past real-world events requires finding meaningful patterns that may not be visible within the complex stream of potentially relevant data currently available on the Web. But Web-based reports of events are often highly redundant and frequently contradictory or ambiguous. Moreover, temporal and spatio-temporal patterns of events tend to have hierarchical structures with interactions and feedback effects over multiple spatial and temporal scales. To address this problem, we continue to develop a sophisticated pattern analysis system, called STempo, which provides integrated computational/statistical and visualization tools to help reveal and explore the complex structure of these dynamic associations.

Within STempo we have implemented a statistical/computational technique based on T-pattern analysis for discovering hierarchical associations among events and groups of events. Using data derived from RSS newsfeeds we demonstrate the effectiveness of our technique by comparing patterns of events in Yemen at different time periods and to events in other Middle Eastern countries during Arab Spring.

2. Background

There have been significant advances in pattern recognition techniques in the field of data mining, and more specifically, in knowledge discovery from database (KDD) techniques within GIScience. Many procedures have been developed explicitly for temporal and spatio-temporal pattern discovery. These are commonly called sequence analysis procedures and are most frequently used to analyze such things as buying behavior and patterns of bank transactions (Gabadinho et al. 2011). There has also been much activity in analyzing point movement patterns (Joh et al. 2002, Dodge et al. 2012). Nevertheless, existing pattern discovery and pattern analysis techniques tend to focus on the temporal interrelationships among a very limited number of event types. To address the need for computational techniques to help reveal distinct but overlaid temporal relationships in complex space-time data involving many event types, we adapted and extended a statistical pattern discovery technique known as T-pattern analysis (Peuquet 2012).

3. Brief Overview of the T-pattern Method

T-pattern analysis, originally proposed by (Magnusson 2000) is designed to detect patterns of associations among event-types, where a specific sequence of event types recurs more often than is likely by chance. Data is organized in a series of timelines, where each timeline consists of the ordered sequence of times when events of a given event type occurred. There is thus one timeline for each event type. This ordering also allows redundancies to be eliminated, as only the first of multiple events with the same type and timestamp will be retained in the analysis. Events may have duration, however. These are represented as a sequence of consecutive time stamps.

In order to detect patterns, two timelines of occurrences for two different event types are compared to find whether the temporal distances between co-occurrences are random or not with respect to a user-specified level of statistical significance (p-value). This test, called the Critical Interval (CI) test, is based on the null hypothesis that pairs of events (e.g., A and B) are temporally independent of each other against the alternative that the real time differences, *d*, between them are relatively invariant. Further, if a type A event occurs at time t_i, the CI test checks whether there is an event type B occurring within a time interval of $[t_i+d_1, t_i+d_2]$ (where $d_2>d_1\geq0$) more often than we would expect by chance. Since there can be no assumption about which event type tends to occur before the other, the CI test checks both BA and AB occurrences. The temporal distance given in the CI, as well as the use of a probabilistic approach accommodates non-stationarity of event patterns and provides a degree of fuzziness in pattern identification.

The pattern detection process proceeds in a bottom-up fashion by comparing each event timeline with all other event timelines, including timelines that had been combined from previously detected groupings of event types. Resulting patterns may be complex hierarchies comprised of multiple event types. T-pattern analysis only identifies as patterns those sequences of event types that consistently recur within a given span of time, along with expected temporal distance (the CI) between each of those events. There is no implication of causality.

4. Validation Analysis Design

Given this probabilistic approach, we would not expect T-pattern analysis to find significant patterns in randomized data. We thus designed a more complex validation analysis to compare results using subsets of real-world data.

Patterns of real-world events change and evolve over time. Also, patterns of events can be expected to recur in places that are similar. Results from T-pattern analysis should therefore reflect this. To test our expectations of how T-pattern analysis will perform on real-world event data, we select a specific pattern discovered to be significant by T-Pattern analysis and test whether this pattern is found in other spatial and temporal contexts. Where we do find recurrence, we examine potential changes in CI values.

While T-pattern analysis can be applied to any domain of real-world events and at any spatio-temporal scale (the spread of disease, climate change, etc.), our validation analysis focuses on events in the Middle East. In particular, the events of Arab Spring embody new social and political patterns that cannot always be assumed based on *a priori* canonical examples. But not all countries in the Middle East have undergone political change. Some remain stable. We therefore expect to find variations in patterns of social and political events between countries and variations over time for a country undergoing change.

4.1 The Data

We captured RSS news feed data and subsequently encoded event types for each data observation from the article titles using the TABARI and CAMEO categorization tools (Schrodt et al. 2008). We found that some built-in CAMEO event types (i.e., ‘reject’ and ‘disapprove.’) are overly vague for the purposes of finding meaningful patterns. We therefore also implemented a post-processing program to expand the event ontology and further specify the frequently generic event types into more specific categories. Because event location is often not mentioned in news report titles, we used part-of-speech tagging on the body of the story to determine geographic location. HTML tags contained within the source code of web-based news reports were used to extract the body of the story.

4.2 Comparison of Patterns among Times and Places

We began by running the T-pattern discovery process on event data for only the country of Yemen from February, 2011 through March 2012 as our basis for comparison. Figure 1 shows one pattern revealed by T-pattern analysis that is seemingly indicative of the socio-political instability in Yemen at that time. Event types are represented by numerical codes, and event transition is given as the CI in number of days.

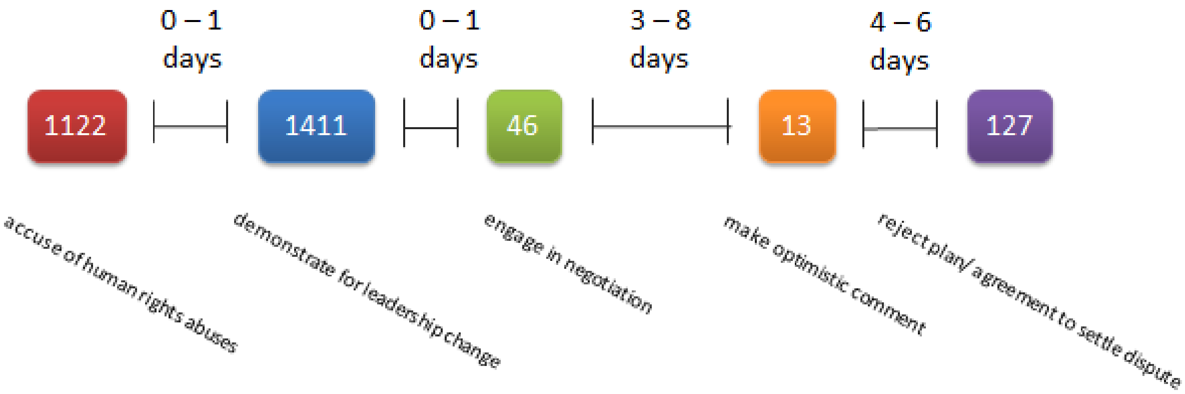
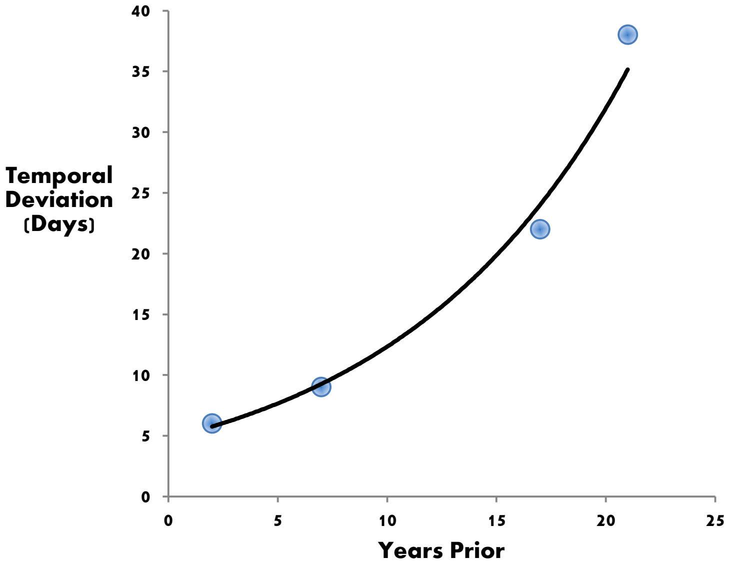

0 – 1 days
0 – 1 days
3 – 8 days
4 – 6 days
1122
1411
46
13
127
accuse of human rights abuses
demonstrate for leadership change
engage in negotiation
make optimistic comment
reject plan/ agreement to settle dispute

Figure 1: A temporal pattern of diplomatic cooperation discovered in Yemen

An accusation of human rights abuses (event code 1122) is followed within one day by a demonstration for a change in leadership (1411), followed within another day by engagement in negotiation (46), followed between three and eight days by an optimistic comment being made (13), concluding four to six days later with rejection of the plan that would have settled the dispute (127).



We then ran additional T-pattern analyses on Yemen for four temporal intervals that were politically unstable times for that country (2009-10, 2004, 1994, and 1989-91). We found this same pattern, but with differing CI values. As shown in Figure 2, CI values (in number of days) increase with increasing temporal distance (in years) from the 2011-2012 temporal interval. This would indicate an evolving situation in that country.

For our continuing validation analysis, we selected Syria, Libya and Egypt as countries with similar socio-political instability, and Bahrain, Jordan and United Arab Emirates as countries that were stable during this period.

REFERENCES

Dodge, S., Laube, P. and Weibel, R., 2012. Movement similarity assessment using symbolic representation of trajectories. *International Journal of Gegraphical Information Science*, 26(9), pp. 1563-1588.

Gabadinho, A., Ritschard, G., Studer, M. and Müller, N.S., 2011. Extracting and rendering representative sequences. In *Knowledge Discovery, Knowledge Engineering and Knowedge Management*, J.L.G.D. A. Fred, K. Liu and J. Filipe (Ed.), pp. 94-106 Berlin: Springer-Verlag.

Joh, C.-H., Arentze, T.A. and Timmermans, H.J.P., 2002. Activity Pattern Similarity: A Multidimensional Sequence Alignment Method. *Transportation Research Part B*, 36, pp. 385-403.

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

Peuquet, D.J., 2012. A method for discovery and analysis of temporal patterns in complex space-time event data. In *Seventh International Conference on Geographic Information Science (GIScience 2012)* Columbus, OH.

Schrodt, P.A., Yilmaz, Ö., Gerner, D.J. and Hermreck, D., 2008. The CAMEO (Conflict and Mediation Event Observations) actor coding framework. In *International Studies Association Annual Meeting* San Francisco.