# Automated Discovery and Modeling of Sequential Patterns Preceding Events of Interest

Kurt Rohloff
*BBN Technologies*
*Cambridge MA, 02138*
*krohloff@bbn.com*

**Abstract.** The integration of emerging data manipulation technologies has enabled a paradigm shift in practitioners' abilities to understand and anticipate events of interest in complex systems. Example events of interest include outbreaks of socio-political violence in nation-states. Rather than relying on human-centric modeling efforts that are limited by the availability of SMEs, automated data processing technologies has enabled the development of innovative automated complex system modeling and predictive analysis technologies. We introduce one such emerging modeling technology – the sequential pattern methodology. We have applied the sequential pattern methodology to automatically identify patterns of observed behavior that precede outbreaks of socio-political violence such as riots, rebellions and coups in nation-states. The sequential pattern methodology is a groundbreaking approach to automated complex system model discovery because it generates easily interpretable patterns based on direct observations of sampled factor data for a deeper understanding of societal behaviors that is tolerant of observation noise and missing data. The discovered patterns are simple to interpret and mimic human's identifications of observed trends in temporal data. Discovered patterns also provide an automated forecasting ability: we discuss an example of using discovered patterns coupled with a rich data environment to forecast various types of socio-political violence in nation-states.

## INTRODUCTION

A major challenge in analyzing complex systems is identifying patterns of behavior which are symptomatic precursors to Events of Interest (EoIs) such as onsets or terminations of socio-political violence in nation-states. By identifying patterns of behavior that precede EoIs such as onsets or terminations of socio-political violence in nation-states we can begin to both understand the underlying causal structures which drive these events to occur and consequently forecast these events.

In this paper we discuss a generalizable sequential pattern concept based on the supposition that the phenomena which cause (or at least are related to) the occurrences of EoIs exhibit similar symptomatic behaviors across multiple EoI occurrences. For example, countries experiencing rebellions driven by the desire for freedom by internal ethnic groups commonly exhibit increasing ethnic tension and violence before the occurrence of ethnic rebellions. We formalize our sequential pattern concept using a finite-state machine model of countries' behaviors and use collections of sampled factor data to define the ``states'' of a complex system such as a country.

The sampled factor data represents quantifiable measurements of systems such as countries at discrete, regular points in time. We use a discrete clock-tick formalism to model the updating of state locations. Example factors from our socio-political domain include GDP, the rates of occurrence of various words in the national press, the average caloric intake, Goldstein measures of conflict/cooperation between governmental entities, etc. These example factors change continuously over time which motivates our use of sampled data. We map the sampled factor data to observed ``trends'' in this factor data where a factor's sampled measurement can be increasing, decreasing, or fluctuating over either the short-term or the long-term.

Although there are numerous published works on pattern discovery, the innovation in our approach to pattern discovery comes from our handling of approximate matches necessitated by noisy data in an application context where rigorous matching may not

always be relevant. In particular, we have developed technology to:

- Define loose matching of observed data trends as part of pattern discovery and matching.
- Numerically optimize pattern matching parameters that are EoI-independent for improved forecasting.
- Discover an algorithm to quickly identify loosely matching patterns

Taken together, these innovations enable our pattern discovery and forecasting approach.

In our motivating context of socio-political violence we are interested in patterns that match the trends of observed behaviors preceding at least two instances of EoI occurrences and which are not present in countries when an EoI does not occur over historical data. We have a generalizable, computationally efficient branch-and-bound back-chaining method to identify the set of factors which define a state space in patterns that match the behavior preceding EoI occurrences in at least two countries from historical data. The backwards chaining methodology permits us to identify which factors change similarly for multiple countries for several time steps leading up to the socio-political violence onset or termination in selected countries in a computationally efficient manner.

As a result of our hypothesis that the phenomena which cause (or at least are related to) the onsets and terminations of socio-political violence exhibit similar symptomatic behaviors across multiple onsets and terminations of socio-political violence, we can generate real-time early-warning forecasts of EoIs if early portions of the patterns are observed in a specific country. This forecasting process is based around the notion of *matching* a country's behavior to early parts of historical patterns. If the country's behavior matches the early parts of the pattern then we forecast that onsets or terminations of socio-political violence will occur in the country in the near future. We found that this approach to forecasting using single patterns is inadequate in practice because individual patterns provide a limited representation of the full breadth of all possible behaviors that may precede onsets and terminations of socio-political violence. This motivates our need to generate libraries of patterns that provide a broader representation of the observed preceding dynamics associated with the occurrence of onsets and terminations of socio-political violence.

We demonstrate our pattern discovery and forecasting methodologies over data of onsets and terminations of ethnic-religious violence in Pacific-region countries from 1998-2006. We show that by discovering patterns for ethnic-religious violence onset and ethnic-religious violence termination over Pacific-region countries from 1998-2004, we can use these patterns to forecast ethnic-religious violence onset and ethnic-religious violence termination over 2005-2006 with a very low false-alarm rate.

Previous versions of our pattern discovery approach is presented in [2],[3]. A more in-depth version of the work presented in this paper is provided in [4]. An introduction to our underlying rich data environment to support the experiments discussed here is provided in [1],[5].

## PATTERNS

We define our patterns to be sequences of trends of behaviors observed in factors before EoIs. We look for trends in sampled factor data where the sampled factors either increase, decrease or fluctuate over either the short-term or the long-term. For instance, in India in the quarters preceding the onset of ethnic-religious violence in early 2002, we see that the level of cooperating expressed by the government towards opposition parties (as measured by a Goldstein metric) holds fluctuating for several quarters before increasing over a short term and then decreasing shortly before the onset of violence.

With this in mind, we formally define the 6 possible types of trends that can be observed in factors as:

- Long-Term Increasing
- Short-Term Increasing
- Long-Term Fluctuating
- Short-Term Fluctuating
- Long-Term Decreasing
- Short-Term Decreasing

Our definition of patterns around these observed factor trends is one of our innovations in our pattern definition. This approach to pattern definition allows for the loose matching of patterns to observed factor data. This loose matching procedure is generally simple for humans, but exceedingly difficult to automate in a computation environment.

For our application context of socio-political violence we define *short-term* trends as those occurring over 3 quarterly time samples or less, and *long-term* trends occurring over 3 quarterly time samples or more. We allow the definitions of increasing, decreasing, and fluctuating to be system- and factor-specific. In our socio-political violence context, the definitions of increasing, decreasing, and fluctuating vary from country to country and factor to factor. Our motivation for this intuition is that ``normal'' observed factor behaviors change differently not only from factor to factor (as may be intuitive because different factors measure different phenomena), but that ``normal'' observed factor behavior varies from country to country for the same factor. As an example, any small change in the level of cooperation expressed by the Chinese government towards potential opposition parties is unusual and significant, but relatively dramatic observed changes in the level of cooperation expressed by the Indian government towards opposition parties is fairly routine.

To map observed changes in factor data to increasing, decreasing or fluctuating trends, we use a weighted threshold test based on the standard deviation of the changes in the factor over a set of training data. This parameter is EoI-independent and is our method for finding this threshold is one of our other innovations in this pattern approach. Our general approach to setting the increasing, decreasing, and fluctuating thresholds is to find the thresholds that would result in maximum forecasting performance over some set of training data.

## FORECASTING

After discovering a set of patterns that precede EoIs in complex systems over some training data, we can use these patterns to make out-of-sample forecasts for the EoIS over test data. We found that a relatively simplistic approach to forecasting is generally very effective - we used a weighting voting mechanism where the discovered patterns matched out-of-sample observations in the test data to generate forecasts.

To implement our weighted voting mechanism in our socio-political violence domain for a given country at a given time, we determine which patterns match the observed factor data leading up to that time. If the number of patterns matching the data exceeds a voting threshold v, then we forecast the onset/termination of socio-political violence in that country at that time. Similar to the weight threshold for increasing, decreasing and fluctuating, we compute v to maximize forecasting performance over some training data that wasn't also used for pattern discovery.

## EXAMPLES OF FORECASTING ONSET AND TERMINATION OF ETHNIC-RELIGIOUS VIOLENCE

Using the trend weight threshold w and the voting threshold v that maximized the f-measure of forecasts for the onset of coups in our training data, we applied our approach to forecast the onset and termination of ethnic-religious violence in Pacific-region countries. We ran this experiment to forecast the onset and termination of ethnic-religious violence using a set of quarterly sampled Goldstein metric factor data that expressed the relative levels of conflict/cooperation between political groups operating in the countries (such as the government, opposition parties, international organizations, etc...).

We split our data into training and test sets. The training data ran from 1998-2004 and the test data ran from 2005-2006. Over the training data there were onsets of ethnic-religious violence in the following countries at the following times:

- China Q1-2004
- India Q1-2002
- Indonesia Q1-1999
- Solomon Islands Q1-2000
- Solomon Islands Q1-2003
- Sri Lanka Q1-2003

Similarly, there were terminations of ethnic-religious violence in the following countries at the following times:

- India Q1-2004
- Solomon Islands Q1-2001
- Solomon Islands Q1-2004

Using these two sets of events and the w = 0.2 threshold, we discovered 55 single-factor patterns for the onset of ethnic-religious violence and 19 single-factor patterns for the termination of ethnic religious violence. When then used these patterns to forecast the onset/termination of ethnic-religious violence using our threshold voting mechanism.

For the onset of ethnic-religious violence, we generated the following forecasts:

- India Q1-2005
- Nepal Q2-2005
- Taiwan Q4-2006

Over the 2005-2006 test data, the only onset of ethnic-religious violence is in India in the beginning of 2005. There are no true ethnic-religious violence outbreaks in Nepal or Taiwan so we generated two false-positive forecasts. It is interesting to note however, that in early to mid-2005 in Nepal there was an uptick in the level of violence associated with the smoldering Maoist insurgency in that country.

For the onset of ethnic-religious violence, we generated the following forecasts:

- China Q1-2005
- Sri Lanka Q1-2006
- Sri Lanka Q3-2006

Over the 2005-2006 test data, the only termination of ethnic-religious violence is in China in the beginning of 2005. There are no true ethnic-religious violence terminations in Sri Lanka so we generated two false-positive forecasts.It is similarly interesting to note however there was a dip in the ongoing ethnic Tamil insurgency in the beginning of 2006 in Sri Lanka, but this violence picked up again several months later.

For both forecasting both the onset and termination of ethnic-religious violence, we were able to forecast both occurrences of onset/termination correctly along with a reasonably low false-positive rate. Because we were forecasting over 29 countries and two years, the false-positive rate is approximately one false alarm every 100 country-quarters.

## A RICH DATA ENVIRONMENT TO SUPPORT PATTERN DISCOVERY AND FORECASTING

In implementing our socio-political violence forecasting methodology, our experimentation environment was supported through the application of emerging information technologies to run our pattern discovery and forecasting methodology. Key technologies in our environment include automated data collection, knowledge representation, model integration and data visualization. To support our pattern discovery and forecasting activities we constructed an end-to-end distributed knowledge system that supports:

- Automated collection and classification of unstructured data (such as raw news feeds and communications data); and collection of structured data.
- Automated fusing of extracted information with historical structural and geospatial datasets using Semantic Web technologies to support distributed modeling and analysis.
- Context-dependent data visualizations, including faceted browsing and spatial-temporal displays, to reveal underlying structures, patterns, and correlations.

Our system utilizes technologies for automated collection and classification of unstructured data including a set of named entity, relationship, and event extraction capabilities that operate over the entire content of articles and can "learn" or evolve over time. We applied these technologies to analyze a voluminous corpus of news feed data that covered a wide geographic region over a period of ten years. The extracted information formed the basis of theory-based independent variables (such as general tension metrics, non-state actor attributes, and leadership characteristics) as well as augmenting the more stale historical factors extracted from existing social science and econometric datasets.

Using Semantic Web technologies, our knowledge system fuses information extracted via the natural language tools

described with data from numerous social science datasets to develop a knowledge environment. This capability also provides a basis for automated reasoning and inference for model and analysis results integration. By directly encoding the semantics typically stored in dataset codebooks, the system fuses multiple datasets that goes beyond superficial dataset alignment (such as merely sorting data from various datasets by county and year). Our system's knowledge infrastructure supports model-agnostic access to the stored data for further manipulation and analysis. Modeling and analysis results can be fed back into the knowledge system for access by other models.

The capstone interface of our knowledge system employs data visualization techniques to display data analysis results and provide interactive "drill-down" capabilities to better study results. Faceted browsing of factors and patterns based on these data values allows a user to select different events and the associated variables associated with the events in various countries.

**DISCUSSION**

The sequential pattern methodology is an approach to identify temporal patterns of behavior that precede EoIs such as multiple occurrences of socio-political violence such as riots, rebellion onset, coups, etc. Our current development of the sequential pattern methodology mines regularly sampled factor data collected from our data collection processes and aggregated by our Semantic Web technology. This methodology identifies commonalities preceding multiple events of interest occurrences but not before non-occurrences. The patterns rely on the pre-defined of equivalence classes of factor values. Our latest experimentation with the sequential pattern methodology relies on equivalence classes defined by both static and dynamic quantization operations.

The sequential pattern methodology is a groundbreaking approach to pattern discovery because it generates easily interpretable patterns based on direct observations of sampled factor data. The resulting patterns are easily to visually interpret as timed finite-state-machine models. Additionally, they permit a clear audit trail to guide the replication of discovered patterns and for the forecast of political instability based on already discovered patterns. We foresee additional development of our sequential pattern technology to natively incorporate discrete-event behaviors to capture the underlying "mode-switching" behavior inherent in many processes.

An important property of our sequential pattern approach is its generalizability. The sequential pattern methodology can be applied in other complex system application contexts where patterns of behavior in sampled factor data preceding events needs to be identified and represented.

# References


1. Battle, R., D. Reid & K. Rohloff. "CWEST: Disruptive Integration of Computation Technology for Data Analysis and Visualization." Visualizing the Past: Tools and Techniques for Understanding Historical Processes, February 2009.
2. Rohloff, K. & A. Victor. "The Identification of Sequential Patterns Preceding the Occurrence of Political Events of Interest." Second International Conference on Computational Cultural Dynamics, September 2008.
3. Rohloff, K. & A. Victor. "Computational Methods to Discover Sets of Patterns of Behaviors that Precede Political Events of Interest." AAAI Spring Symposium on Technosocial Predictive Analytics, March 2009.
4. Rohloff, K. "Trend Pattern Approach to Forecasting Socio-Political Violence." Third International Conference on Computational Cultural Dynamics, September 2009.
5. Rohloff K. and W. Thornton. "A Knowledge Environment for Social Science Exploration." Human Behavior-Computational Intelligence Modeling Conference, June 2009.