Events, patterns and analysis: forecasting conflict in the 21st century

Devika Subramanian and Richard Stoll

April 2004

1 Introduction

With the explosion of online news over the past few years and with recent advances in extracting information from text, the question we study is: is it possible to forecast the outbreak of serious conflict by monitoring news media from regions all over the world over extended periods of time, extracting information about events from them, and computationally analyzing the time series of extracted events? Our research is motivated by this question and this abstract presents preliminary studies that suggest the potential of automated events data extraction and analysis for predicting and understanding conflict.

Specifically, the goals of our research are:

1. To design information extraction techniques and to build events data sets from online news stories, for use by the entire scientific community.

2. To develop the algorithmic base for making predictions about the onset of serious conflict from events data.

Timely warning of the outbreak of serious conflict can be a key element in conflict resolution. Early warning can provide the time for state and non-state actors to intervene and prevent the outbreak. Thus, we feel our work can be of potential value to the conflict resolution process, even though the focus of our research is the outbreak and evolution of conflict.

2 The raw material: Gathering news stories

A key issue is the selection of online media outlets to be used as the source of coding of events data. We have focused on the following media sources based on their electronic availability and their reputation: Washington Post, BBC, Agence France Presse (AFP), Reuters, Boston Globe, Houston Chronicle, Scripps Howard Newswire, and the Associated Press. We designed a suite of programs to extract relevant news stories from the websites of these news sources. As of this report filing, we have obtained all articles pertaining to the Middle East from 1979 to 2003 of the Boston Globe, from 1985 to 2003 of the Houston Chronicle, from 1990 to 2003 of the Scripps Howard newswire, from 1977 to 2003 of the Washington Post, and from 1998 to 2003 of Associated Press and the BBC. The news articles are stored in a custom relational database. Such a representation facilitates complex queries over the stories and allows efficient access to them for further analysis. The database also offers support to link events data with the news stories they are generated from. This database will be a useful tool for extracting and coding a diverse set of events data collections.

3 Filtering irrelevant stories

To our surprise, only 10% of the stories retrieved from the news sources above contain codeable events. The majority of the stories do not pertain to conflict, or are not factual reports of conflictual events (e.g., are opinion pieces). Effectively filtering these stories is a key step in the analysis. We have trained probabilistic classifiers to identify codeable stories from uncodeable ones. The classifier's accuracy is 90% and we are working to improve the reduce its false positive and negative rates even further.

4 The Rice event data extractor

We extract events from each codeable news story using a class of programs called event extractors — these programs analyze the text of the story (including the headline) and gather four pieces of information: who did what to whom and when. The actors are countries or organizations (e.g, Hamas or LET). The events are coded in a conceptual hierarchy called WEIS, developed by political scientists and used by them for manual coding of news stories. Each event type has a scaled score (-10 (conflict) to +10 (cooperation)) associated with it.

There are lots of computational problems in extracting events from news stories. Identification of actors and events in non-trivial, and involves named-entity recognition. We use statistical parsing, and part of speech tagging and learn probabilistic recognizers of actors and targets. In a similar fashion we learn models that map verb phrases into the event categories in the WEIS hierarchy.

5 Analyzing events data

We analyzed a previously gathered event data collection coded from Reuters stories from eight countries in the Gulf region of the Middle East for the period 1979 to 1999. We aggregated the scaled event scores for these eight countries on a biweekly basis. Each point in the time series represents the average event scores for a 2 week period. This series is highly non-stationary, so traditional methods of prediction do not perform well. We use wavelets to do a multiscale decomposition of the signal, and find discontinuities in the series. These discontinuities occur across all scales and are identified using Mallat's modulus maximum technique. Surprisingly, these discontinuities correspond almost exactly to major conflicts in the region! These include the start and end of the Iran/Iraq war, the Gulf War and the first and second Intifada. Further, the wavelet coefficients preceding a singularity or discontinuity exhibit trends that allow us to predict conflicts with high probability about 8 weeks before actual hostilities break out. We have repeated our analysis technique on Cold War events data over a period of fifteen years and developed a computational chronology of key events in that region. The detailed results are reported in an accompanying Powerpoint presentation.

A lot remains to be done. We are working to refine our story filter, event extractors, and time series analyzers. We are gathering information about events that led up to the 2003 Iraq war to provide a computational account of history based on extracted events alone. We are also experimenting with the use of novel statistical methods for analyzing data and considering ways of introducing more information (e.g., economic factors) into our analysis.