CWEST: Disruptive Integration of Computation Technology for Data Analysis and Visualization

Robert Battle, Douglas Reid, Kurt Rohloff
BBN Technologies
10 Moulton St.
Cambridge, MA 02138
{rbattle,dreid,krohloff}@bbn.com

The integration of emerging information technologies from data collection to analysis and visualization via distributed computation technologies has enabled a paradigm shift in researchers' abilities to understand historical events and processes as they spread out across both time and space. We propose to present a seminar on our ongoing work, lessons learned and research vision for several technologies we are developing and integrating as part of the CWEST information system at BBN Technologies.

The CWEST information system is an automated system that collects and processes structured and unstructured data (such as raw newsfeed and communications data), fuses the extracted information with historical structural and geospatial datasets and reference information, and then analyzes the resultant information to identify patterns that are used to forecast political instability events such as coups, the onsets of rebellions, and international crises. The system exploits state-of-the-art technologies in natural language processing, data fusion, symbolic processing, and pattern recognition.

We propose to discuss both our internally developed component technologies as well as the overall CWEST system that leverages the benefits of the component technologies to enable deep insight into historical process across the spatial and temporal domains. In particular, the technologies we will discuss include:

- Information extraction technologies for data collection and dataset generation
- Semantic Web technologies enabling fusion of large-scale, diverse datasets
- Sequential pattern methodologies to discover and analyze temporal patterns of behavior exhibited by countries before political instability
- Context-dependent visualizations, including faceted browsing and spatiotemporal displays, to expose both structure and impact of data and resultant analyses.

Our information extraction tool suite includes a set of named entity, relationship, and event extraction capabilities that operate over the entire content of articles. For the CWEST project, we applied these technologies to a voluminous corpus of newsfeed data covering a period of ten years for a wide geographic region. The extracted information formed the basis of theory-based factors for analysis (general tension metrics, non-state actor attributes, and leadership characteristics) as well as augmenting stale historical factors extracted from existing datasets.

As part of our long-term vision for information extraction, we envision enhancing our current capabilities by focusing on extraction tailored to the needs of social science and humanities-driven historical analyses. Our CWEST experience revealed that the

development of novel capabilities, such as identifying speaker text, linking actors to their political ideologies, event de-duplication and enhanced extraction of temporal and spatial properties for events, provides significant analytic benefits. Primarily, these developments will enable us to generate more accurate factors to be used for our sequential pattern analysis process. The development of these capabilities will also enable the inclusion of new factors in the analytic process as well as potentially enabling new analytic vectors when considering historical cases.

The Semantic Web provides a common framework for integrating data from many disparate sources in a rich manner. One of the largest benefits of our use of Semantic Web technology is that its encoding provides a basis for automated reasoning and inference. For the CWEST effort we have been using Semantic Web technologies to fuse data from numerous tabular social science datasets with the information extracted via the natural language tools described above into a knowledge base. By encoding the semantics typically stored in codebooks directly in an exploitable manner, we can achieve a data fusion that goes beyond superficial dataset alignment based on properties such as county-code and year. This enables the generation of new factor data for consideration that eclipses the original scope of the dataset and/or that could only have resulted from the combination of multiple datasets. As a result, we are able to consider factors for analysis that were not historically available, as well as seamlessly augment existing and generated datasets with geospatial and temporal extents.

Part of the vision of our use of Semantic Web technologies is the deployment of further ontological modeling of the source data that will enable more advanced reasoning over the data corpus and allow for richer analytic constructs. One potentially rich avenue for development is anomaly detection based on ontological similarity, which we believe to be vital for historical pattern development. Additionally, by representing what a factor "means", we hope to more fully de-couple the analytic process from the actual data sources feeding it. This should allow for the development of more broadly applicable patterns, which can be run against a variety of data repositories.

The sequential pattern methodology is an approach to identify temporal patterns of behavior that precede multiple occurrences of political instability events such as riots, rebellion onset and coups. 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.

As part of the capstone interface of our CWEST system, we employ data visualization techniques that allow us to display patterns and provide interactive drill-down for audit trail to verify discovered patterns. Importantly, this technology allows us to display a discrete-event system representation of our sequential patterns. We can drill-down into the discrete states of the sequential patterns to identify the factor data values which define the state and the specific countries which follow the pattern at various historical instances. From this level of pattern drill-down view of the factor values we display the geographic distribution of actions that are following the sequential pattern. Faceted browsing of factors and patterns based on those factors allows the user to select which aspect of the pattern and factors are viewed during the inspection of historical cases.

Our vision of our ongoing development of data visualization includes the implementation of analyst-driven "what-if" analysis, via online manipulation of datasets and analytical tools, to examine the impact of specific factors on historical scenarios. Additionally, we recognize the need for the development of analysis-unit-specific views of the information that present both context and high-value information in a unified, succinct manner. In combination, these capabilities will enable analysts to better understand the patterns generated by our automated system and to rapidly test novel explanatory theories.

We propose to discuss the technologies described above in the context of conducting historical analyses with the CWEST system. Through its component technologies that collect, integrate, process, and display patterns of behavior present in historical cases of political instability, the CWEST system enables the investigation of historical commonalities in temporally and geographically distributed behaviors.