Predictive Analytics Using a Blackboard-based Reasoning Agent

Jia Yue[1], Anita Raja[1], William Ribarsky[2]
[1]*Department of Software and Information Systems*
[2]*Department of Computer Science*
The University of North Carolina at Charlotte
{jyue,anraja,ribarsky}@uncc.edu

Abstract

Significant increase in collected data for investigative tasks and the increased complexity of the reasoning process itself have made investigative analytical tasks more challenging. These tasks are time critical and typically involve identifying and tracking multiple hypotheses; gathering evidence to validate the correct hypotheses and eliminating the incorrect ones. In this paper we specifically address predictive tasks that are concerned with predicting future trends. We describe RESIN, an AI blackboard-based agent that leverages interactive visualizations and mixed-initiative problem solving to enable analysts to explore and pre-process large amounts of data in order to perform predictive analytics. Our empirical evaluation discusses the advantages and challenges of predictive analytics in a complex domain like intelligence analysis.

1. Introduction

Consider the scenario where an intelligence analyst is informed that a terrorism event occurred several minutes ago. The analyst is provided with partial and sometimes conflicting information about the incident including the incident location, the occurrence time, the attack weapon, and the number of human casualities reported in the various news media. The analyst has to determine the terrorist group/organization behind the incident, which is a critical piece of information related to the entire attack. Once the name of the terrorist group/organization has been identified, the activity trend of the group is determined so that appropriate security and response measures can be taken. Data on past terrorism events is used to uncover the group's terrorist activities by year and region. This will provide the analyst with a high-level picture of the activity for this group. For example, a certain group A may have the activity pattern of launching a major attack and then not being active for the next six months. On the other hand, a rival group B, which is also active in the region, could have the activity pattern of initiating several small transportation related attacks simultaneously or in spurts over a period of 2 or 3 days. Based on the characteristics of the current event and past activity patterns, the analyst can determine the trend of the future events and suggest appropriate precautionary measures.

Pirolli and Card [8] describe a model of cognitive task analysis performed by analysts that consists of two overlapping loops. The first of the two loops is the foraging loop that includes searching for information by accessing external data sources; reading and extracting the data; and searching and filtering evidence to support the second loop called the sensemaking loop. Sensemaking involves hypothesis generation and validation and uses the evidence generated by the foraging loop to support its search process.

In this work, we are interested in building an automated mixed-initiative agent called RESIN (RESource bounded INformation gathering) [6] to assist analysts in their decision-making. This agent will help determine predictions of immediate and near-term events based on past events by seamlessly integrating visualizations and predictive analytics to support the foraging and sensemaking loops. Predictive analytics is concerned with the prediction of probabilities of future events and trends based on observed events. It encompasses a multi-perspective approach that includes integrated reasoning, information visualization, pattern recognition and predictive modeling associated with domain knowledge. Recent progress in information visualization integrates new computational and theory-based tools with innovative interactive techniques and visual representations to enable richer access to data.

RESIN extends our previous work on TIBOR [1], a foraging agent capable of making time-bounded decisions. It combines both foraging and sensemaking loops by emphasizing the blackboard reasoning and mixed-initiative reasoning aspects in order to assist investigative analysts in performing visualization-based predictive analytics. It also leverages sequential decision making [2] to determine the sources for foraging analysis and the AI blackboard system [3] to support hypothesis tracking and validation (sensemaking loop) in a highly uncertain environment. Providing a clear explanation in support of such a decision making process is critical, since it is the key to gain and maintain the analyst's trust in the system. Moreover, RESIN provides ways for the user to interact with and control the problem-solving process at all critical decision points defined by the expert. In summary, RESIN provides investigative analysts with the capability to forecast real data; provide access to automated support for their decision-making; find non-myopic alternate solution paths; to investigate outliers in the data. One assumption we have made in this work based on our discussions with intelligence analysts is that past information about terrorist attacks can be useful to forecast/understand future ones.

2. Predictive Analytics in RESIN

The hypothesis we plan to validate in this paper is: Given a current event, it is possible to predict future events by first determining the missing information about the current event and then determining the event trends based on the historical events captured in the global terrorism database.

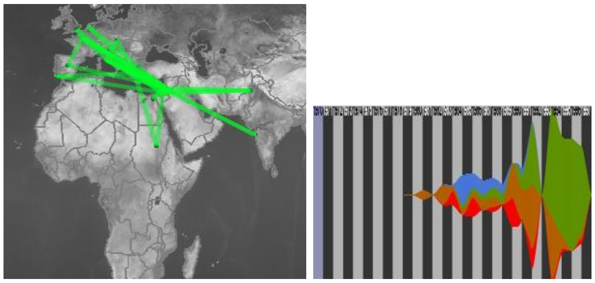

Figure 1. MapView (Left) and Temporal View (Right) for group Hezbollah

The Global Terrorism Database (GTD) [5] is an open-source database containing information on both domestic and international terrorists from 1970 through 1997 (with recent update to 2007). The GTD tool [9] is a visual analytics approach that provides a comprehensible presentation of this massive geopolitical event database. With its four highly coordinated views (corresponding to Who, What, When, Where), this tool will visually provide investigators knowledge about terrorist activities and their relationships and try to guide them to understand why those activities happened through user interactions. We employed the Box-Jenkins [10] procedure, a widely used and efficient forecasting technique, especially for time series, for the analysis of terrorist events from the GTD.

In Figure 1, the *MapView* shows related incidents Hezbollah. The user can use this view to compare group activities and determine for instance that despite sharing similar geo-spatial distributions, these two groups are quite with respect to radius and ranges, suggesting unique terrorist activities. Similarly a *TemporalView* reveals the active times for Hezbollah (1979-1993) and their main attack types. (The colors in the temporal views are keyed to different attack types) The information from these two views will provide the human user with clues to estimate their confidence values and to adjust the predicted results.

RESIN's problem solving process is initiated when the human user identifies a goal that is posted on the blackboard, and this action triggers the RESIN agent. Given a particular deadline for the problem solving process, RESIN will assist the analyst in predicting future events by incrementally identifying the group perpetrating the current event and the behavior patterns of that group so that necessary steps can be taken to prevent casualities and major damage in any possible near-future incidents.

TYPE: Assassination
WEAPON: Explosives
ENTITY: Political Party
YEAR: 1992
REGION: Middle East/North Africa
NKILL: 2
GNAME: ?

Figure 2. Partial Terrorist Incident Description

RESIN's input tuple contains partial information about a single current terrorist incident (see example in Figure 2) and a deadline to solve the problem.

The input event has six known categories: TYPE, WEAPON, ENTITY, YEAR, REGION, and NKILL as initial inputs. Each category has a different number of possible values, for example, TYPE (e.g. assassination, bombing, facility attack) contains different types of attacking methods, while ENTITY represents different attack targets, such as 'Political Party', 'US Police/Military' and so on.

We have identified two phases in the sensemaking loop for the intelligence analysis domain.

Phase1: Category Prediction (CP): Predict the missing group name (GNAME) and associated confidence value of a partially defined current event.

Phase2: Event Count Prediction (ECP): Predict number of events (ENUM) possibly perpetrated by the group identified in Phase1 and confidence value by applying time series analysis on historical data.

The details of each of the phases are described below:

Phase1: CP

[Input]: Input event and all incident data from the GTD

1: Input event posted on the blackboard.

2: TÆMS task structure [4] modeler generates appropriate task structure and translates it to an MDP. The MDP is solved by the MDP solver and suggests C4.5 as the first KS to be used.

3: C4.5 uses the GTD data set and the input tuple as input and generates an initial prediction of GNAME.

4: User verifies the predicted result using GTD tool.

5: The final solution of the predicted group name with a confidence value posted on the blackboard.

[Output]: Predicted GNAME with confidence value is posted on blackboard

Phase2: ECP

[Input]: Input event, Prediction time period, GNAME with confidence value, all incident data from the GTD

1: Manually identify if the data is time series data using log transformation of the initial data variables and a unit root test.

2: Choose appropriate Box-Jenkins method model.

3: Predict ENUM using Box-Jenkins method for given Prediction time period.

[Output]: ENUM with confidence value for given year.

Figure 3. RESIN's 2-phase Predictive Analytics

3. Experiments

We now describe experiments to predict events and their types following the given input incident by employing time series analysis. Here, we employ the Box-Jenkins [7, 10] procedure, a widely used and most efficient forecasting technique, especially for time series, for the analysis of terrorist events in the GTD [5]. We carry out the time series analysis for the prediction from two perspectives: one is the perspective of all available data; the other is focused on the data in a certain region or carried out by a particular terrorism group.

All terrorist events grouped by month of number of events (data in 1993 is unavailable in the GTD) were used for this analysis. There are 324 months' data in the GTD from January 1970 to December 1997 and we employed the first 312 months' data as the training set to create the model while the last 12 months' data as the testing set for the purpose of comparison with the predicted values. That means 12 data points needed to be predicted by the model created by 312 data points. With the log transformation of the initial data variables and a unit root test, we obtained the series which satisfy prerequisites of the Box-Jenkins method that requires the time series to be stationary with constant mean value. With the comparison of the optional models based on three pivotal main parameters-Adjusted R-squared (the coefficient of determination), Akaike Information Criterion (AIC) and Hannan-Quinn criterion ,we determined AR (3) to be the best model for the GTD data series with a Mean Absolute Percent Error (MAPE) index of 15.51916 for static prediction.

Figure 4 indicates the close fit between the predicted number of terrorist events and the actual number of occurrences for the year 1997. In fact, most of the 12 actual values are within the 95 percent confidence interval of the predicted values. Compared with GTD's complete data as shown in Figure 5, our forecasting is on par with the actual curve which suggests that the approach we used here for the prediction with the overall GTD is feasible.

We now define Region data as terrorism event incidents in the past for a particular region while Group data is the set of incident perpetrated by a group. In the GTD, there are six regions and the worldwide attacks are carried out by 2404 terrorist groups. The stationary series for Region and Group were obtained by the transformation of one-order differencing to the initial data. The correlogram analysis indicates that these series are distributed randomly within the confidence interval, which indicates that they are purely stochastic series that cannot fit into AR, MA, and ARMA models [10]. One general method for predicting a random time series is to acquire the mean of its stationary status and the predicted value can be obtained by the converse transformation of differencing [10].

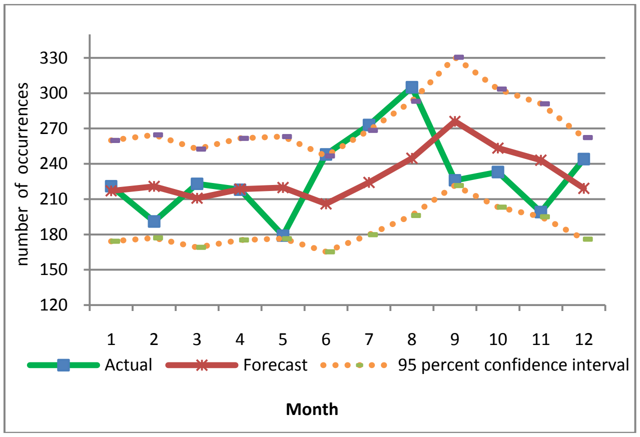

Figure 4. Actual and predicted number of events per month from 1997.1 to 1997.12

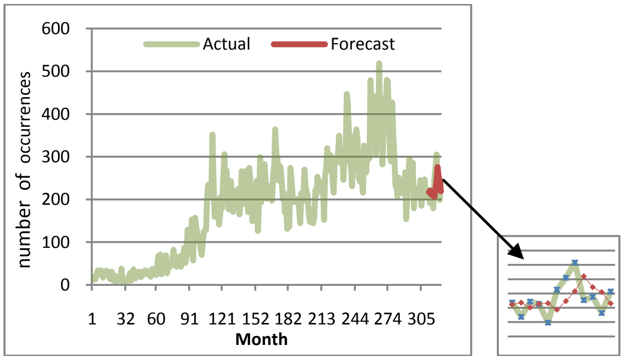

Figure 5. Actual numbers of events from month 1th to 324th with comparison of predicted ones from month 313th to 324th

Table 1 describes the results of the forecasting process on the Region data. There are only 27 years of GTD data available due to the missing data in 1993. It is observed that the large standard deviation (Std Dev) for each Region indicates that the predicted value deviates significantly from the actual value. This is valid since the Region data is quite sparse. Since there are more data points available at the Group level, the observed standard deviations shown in Table 2 are much lower than the Region data. The predicted values are in the same order of magnitude as the actual values. This is encouraging given that our data was sparse. Since our goal in this paper is to identify the appropriate approach for predictive analytics in this domain, we contend that more accurate data would improve our performance. It should also be noted that our approach the predicted value is very dependent on the previous value. This implies that the approach is most effective in domains that require short-term forecasting, such as terrorist event predictions.

Table1. Numbers of terrorist events for Region Data

REGION	Mean	Std Dev	Predicted Time	Predicted Value	Actual Value
North America	-2.00	31.38	1997	17	31
Europe	20.34	159.88	1997	755.34	514
Middle East/ North Africa	13.88	227.56	1997	194.88	554

Table2. Numbers of terrorist events for Group Data

GROUP	Mean	Std Dev	Predicted Time	Predicted Value	Actual Value
Fatah	0.015	1.68	1992.Q1	1.01	1
Hezbollah	0.05	3.89	1997.Q1	2.05	4
ETA	0.058	11.50	1997.Q1	0	10
IRA	-0.01	13.64	1997.Q1	3.98	7

4. Discussion and Future Work

Most of the studies on global terrorism mainly focus on the descriptive analysis of past trends and the exploration of the various causes of terrorism. The predictive analytics on the Global Terrorism Database involves the identification of global patterns and trends as well as high-level strategic reasoning. The RESIN system combines classification techniques, blackboard-based reasoning and the GTD visualization tool to perform the prediction for unknown event information and the future trends. The goal of the empirical study in the previous section is to verify whether our approach is on the right track and how well it performs in a real-world application. A huge challenge in intelligence analysis is the uncertainty (missing and contradictory) information about the incidents. So the quality of our results is not as high as it would be in a simulated domain. However, the results do support the hypothesis that the combination of methods in RESIN will potentially assist in predictive analytics.

In that sense, RESIN is a good start. There are some interesting areas that we would like to investigate in the future. We intend to integrate the AI blackboard with a variety of other knowledge sources, including a sensemaking knowledge source that uses case-based reasoning and pattern recognition; as well as other multimedia visualization tools such as the semantic video browser. We plan to extend RESIN's functionality so that it will facilitate an analyst's problem solving process by determining predictions about an event from multiple and conflicting viewpoints from different stake holders. We want to combine information from multiple sources and using multiple predictive methods to perform the forecasting.

5. Acknowledgement

This work was sponsored by the National Visualization and Analytics Center (NVAC) under the auspices of the Southeastern Regional Visualization and Analytics Center. NVAC is a U.S. Department of Homeland Security Program led by Pacific Northwest National Laboratory. We thank the reviewers for their helpful comments.

References

[1] Liu, D., Raja, A., and Vaidyanath, J., *TIBOR: A Resource-bounded Information Foraging Agent for Visual Analytics*, Proceedings of 2007 IEEE/ WIC/ ACM International Conference on Intelligent Agent Technology (IAT 2007), Silicon Valley, CA, November 2007, pp. 349-355.

[2] Bertesekas, D. and Tsitsiklis, J., *Neuro-Dynamic Programming*, Athena scientific, Belmont, MA, 2006.

[3] Corkill, D., *Blackboard Systems* AI Expert, 1991, 6(9):pp 40-47.

[4] Decker, K. and Lesser, V., *Quantitative modeling of complex environments*, International Journal of Intelligent Systems in Accounting, Finance, and Management, December 1993, 2(4): pp.215-234.

[5] LaFree, G. and Dugan, L., *Global Terrorism Database, 1970 – 1997*. ICPSR04586-v1, College Park, MD: University of Maryland, 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], Apl, 2007.

[6] Liu, D., Yue, J., Wang,X., Raja, A., Ribarsky, W., *The Role of Blackboard-based Reasoning and Visual Analytics in RESIN'S Predictive Analysis*, Proceedings of 2008 IEEE/ WIC/ ACM International Conference on Intelligent Agent Technology (IAT 2008), Sydney, Dec 9-12, 2008.

[7] Brockwell, P., and Davis, R., *Time series theory and methods*, published by Springer, 1991.

[8] Pirolli, P., and Card, S., *The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis*, Proceedings of 2005 International Conference on Intelligence Analysis, 2005.

[9] Wang, X., Miller, E., Smarick, K., Ribarsky, W., and Chang, R., *Investigative Visual Analysis of Global Terrorism,* Journal of Computer Graphics Forum, Eurovis, 2008.

[10] Oppenheim, R.,*Forecasting via the Box-Jenkins Method*, Journal of the Academy of Marketing Science, Vol.6, No.3, June 1978, pp 206-221