A Blackboard-based Approach towards Predictive Analytics

Jia Yue[1], Anita Raja[1], Dingxiang Liu[2], Xiaoyu Wang[2], William Ribarsky[2]

[1]Department of Software and Information Systems
[2]Department of Computer Science
The University of North Carolina at Charlotte
{ jyue, anraja , dliu, xwang25 , ribarsky }@uncc.edu

Abstract

Significant increase in collected data for analysis 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 to eliminate 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 visualization and mixed-initiative problem solving to enable analysts to explore and pre-process large amounts of data in order to perform predictive analysis.

1. Introduction

Predictive Analysis is concerned with the prediction of future probabilities and trends based on observed events. It encompasses a multi-perspective approach that includes integrated reasoning, pattern recognition and predictive modeling associated with domain knowledge. We are interested in building an automated reasoning agent, which will determine predictions based on events from the past. For example, the prediction could involve determining missing or unknown information in current data or the occurrences of some potential events in the near future (next day, two days, one month…).

We have developed RESIN, a RESource bounded INformation gathering agent for visual analytics that extends our previous work on TIBOR [1]. It emphasizes the blackboard reasoning and mixed-initiative reasoning aspects of our agent architecture that will assist investigative analysts in performing viewpoint-based predictive analysis. RESIN leverages sequential decision making [2] and an AI blackboard system [3] to support hypothesis tracking and validation 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. We use an AI blackboard to achieve this goal, which maintains a clear evidential path for supporting and contradicting information while allowing for explicit modeling of concurrent top-down and bottom-up processing [15]. RESIN has the capability to pass information between analysts and itself during the problem-solving process by leveraging an interactive visual analytics tool. Moreover, RESIN provides ways for the user to interact with its problem-solving process or even control it at every step through a rich user interface. By using RESIN, investigative analysts can have access to automated support for their decision making; the capability for finding non-myopic alternate solution paths; and a tool to investigate outliers. In addition, RESIN can assist investigative analysts in performing forecast by providing the predicted missing information and the possible future trends with time series analysis.

In addition to the AI blackboard, RESIN consists of a TÆMS [4] task structure library, a Markov Decision Process (MDP) [2] solver and heterogeneous knowledge sources (KSs) [3]. The AI blackboard contains reasoning results from processing existing information, which includes raw data, various problem-solving states, partial solutions and current goals; the TÆMS is an abstraction of the low-level execution model and captures uncertainty in outcome distributions, while the MDP solver is a probabilistic model, which captures the essence of sequential processes and is used to compute optimal policies that identify, track, and plan to resolve confidence values associated with blackboard objects. The KSs are independent specialist computational modules that contain the domain knowledge needed to solve a problem. The control flow of this predictive analysis process involves handling several issues: choosing the appropriate set of databases, analyzing the high dimensional data; generating one type of decision trees to extract and represent the data; determining appropriate interactive visualization tools for these data; performing reasoning processes; and generating final solutions.

In this paper, we apply RESIN to the Global Terrorism Database (GTD) [5] to perform predictive analysis tasks that determine the missing information about a single terrorism event as well as predicting the probabilities of similar such events in the near future. Using machine-learning classification techniques (as discussed in Section 2), blackboard-based reasoning, the GTD Visualization Tool [6] and the technology of time series analysis, we are able to make predictions based on existing historical data.

2. RESIN's Predictive Analysis

In this section, we provide a description of the prediction process, which is used to determine which terrorist group is likely to be responsible for a particular incident. Our overall approach to this problem is utilizing all the appropriate KSs based on resource constraints and task deadline to match the input event to past events for the prediction of group name.

The AI blackboard, KSs and the control mechanism are three main components of RESIN's architecture [8]. RESIN employs several KSs, including the C4.5 [7] algorithm, the GTD, and an investigative visual analytics system built on the GTD. At appropriate times defined by RESIN's reasoning process, the knowledge source takes relevant information from the blackboard and makes a contribution towards the problem solving with its specialized domain knowledge.

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

Figure 1. Partial Terrorist Incident Description

The problem solving process is initiated when the Human User posts a goal on RESIN's AI blackboard and this action triggers the RESIN agent (Step 1). In this paper, the goal is to predict the missing group name (GNAME) and its associated activity trends within a given deadline based on an input tuple that contains partial information about a single current terrorist incident (Figure 1). The input event as described in Figure 1 has six 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. In Step 2, the TÆMS task structure modeler generates an appropriate task structure and translates it to the MDP solver for action assessment. Using dynamic programming, the MDP solver computes the optimal policy based on resources constraints (e.g. deadline) and determines the best action, which will trigger appropriate methods to perform predictive analysis (Step 3) [1]. Through a built-in user interface, the RESIN agent enables the user to interact with the visual analytics tools supporting the mixed-initiative problem solving process, to validate the initial RESIN results and to post results back to the AI blackboard (Step 4). These initial results could include several hypotheses that are possible solutions to the problem. Using these visualization results as well as previous analysis results, the blackboard will then propagate the evidence information, verify a specific hypothesis with an associated confidence value and generate the predicted group name as final solution.

The control flow described above allows RESIN to largely enhance the accuracy of results in solving automated prediction problems. RESIN also can enable the user to perform those tasks manually through the integration of visual analytics tools. Hence, while taking suggestions from RESIN, the user has the power to revise or even dispute those suggestions. All these are designed to help gain and maintain users' trust and assist them to perform better analytical predictive tasks.

3. A Motivating Example

In this section we present a detailed description of the RESIN agent's reasoning process through a prediction scenario, with the goal of determining the unknown group name (GNAME) based on the current event information in Figure 1 as input and the associated activity trends of that group based on entire GTD. To achieve the goal, we selected 100 historical terrorism incidents as the training set.

Based on the input tuple, the TÆMS task structure modeler generates a task structure that models problem-solving patterns. The top-level task is *Predictive-Analysis*, which is decomposed into two subtasks, *Classification-Algorithm* and *Visualization-Analysis*. The *Classification-Algorithm* will determine the data classification algorithm and *Visualization-Analysis* will trigger the appropriate data visualization tools. To justify the importance of user interaction in a mixed-initiative agent, the RESIN's task structure also provides user interaction options at critical points, such as *Map-View-Interaction-Option* and *Temporal-View-Interaction-Option*. Each of these methods is characterized by quality and duration distributions [1]. The task structure is then translated into a MDP solver by computing the optimal policy. In this example, the policy triggers the C4.5 [7] knowledge source to generate the decision tree and predict the group name. The C4.5 KS can not only generate the view of this decision tree, but also show the specific confidence value of the predicted group name and other alternative group names. The automatic processing of this knowledge source will provide a partial solution, which is posted onto the blackboard. For this example, the C4.5 KS predicts that the group name may be Fatah with confidence value of 0.75, along with one alternative solution that the group name may be Hezbollah, with confidence value of 0.25. Fatah is posted as C4.5's partial solution onto the blackboard due to its larger confidence value.

The MDP generated policy then triggers another knowledge source, the GTD tool, to facilitate the C4.5 KS results. When invoked, the GTD tool first would receive

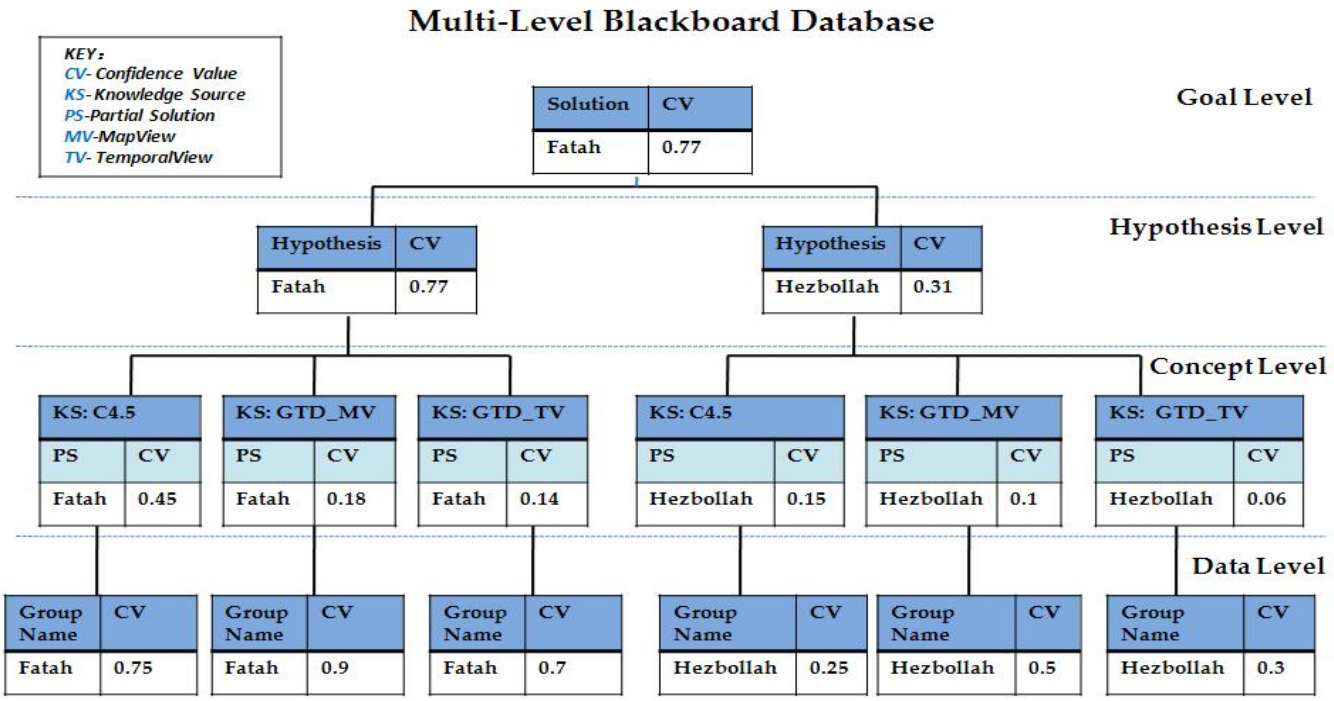

Figure 2. Multi-Level Blackboard Database

relevant information from the blackboard to understand the current status of the problem-solving process and keep track of latest developments of that process. Then through user interaction, the GTD tool will provide detailed information on the input tuples, both in *MapView* and *TemporalView*. With user interaction, the GTD tool posts those confidence values back to the blackboard and updates current information. The combination of all evidence values will be the contribution to the final solution in the Goal level of the blackboard. The group name with the highest confidence value will be posted to the goal level of the blackboard as final solution.

In this example, we compute the confidence value as:

$$GC_{groupname} = \sum_{i=1}^{N} KS_i * KSWt \qquad (1)$$

where N is the number of knowledge sources in RESIN , KS_i is ith knowledge source and *KSWt* is importance weight associated with that knowledge source.

We have three knowledge sources in this example. So the equation is:

$$GC_{groupname} = DT * W_{DT} + MV * W_{MV} + TV * W_{TV} \qquad (2)$$

where DT is the confidence value from the data classification analysis (C4.5 KS); MV is the confidence value from the *MapView* analysis; TV is the confidence value from the *TemporalView* analysis; W_{DT} is the weight of data classification analysis (C4.5 KS); W_{MV} is the weight of *MapView* analysis; and W_{TV} is the weight of *TemporalView* analysis. Through interactions with *MapView* and *TemporalView*, users could determine their confidence values based on the visual patterns shown in these two views. For example, as suggested by *TemporalView*, Fatah has a larger amount of assassination activities in 1992 compared to Hezbollah; the user would agree with C4.5's prediction by providing a high confidence value (0.7). The multi-level blackboard database for the example is shown in Figure 2. It contains four different levels, Goal, Hypothesis, Concept and Data, in order of decreasing granularity. The Goal level stores the goal of the problem and resolution information. The Hypothesis level contains concepts which are represented in the Concept level. The Data level contains the data/evidence gathered to (in) validate the various hypotheses. Assuming W_{DT} = 0.6, W_{MV} = 0.2, and W_{TV} = 0.2, the values of 0.45 (0.75*0.6), 0.18 (0.9*0.2), 0.14 (0.7*0.2) for group Fatah are posted on the Concept level of the blackboard associated with the KS of C4.5, *MapView*, and *TemporalView* respectively. By applying equation (2), we obtain $GC_{Hezbollah}$ = 0.31 and GC_{Fatah} = 0.77 for the Hypothesis level.

Therefore, RESIN predicts the group name is Fatah with the confidence value of 0.77. We then performed time series analysis as described in Section 4.2 to determine the activity trends in the region by Fatah in the near future.

4. Experiments

In this section, we describe experiments to assess the effectiveness of RESIN's blackboard-based reasoning mechanism and to explore the potential of RESIN's

predictive ability. By applying the GTD, group name prediction will determine unknown group name (GNAME) based on partial information of input tuples, while event prediction will address the occurrence of terrorist attack by this group in the following time period.

4.1. Group Name Prediction

This experiment is based on a training set of 2700 incidents selected from the GTD and for each task we use the same ten incidents from a test set, with different deadlines from 30 to 70 (ranging from a very tight to a loose deadline). There are ten users involved in the experiments with access to the GTD tool. Each user will determine the confidence values towards initial predictions with values from -0.9 (strongly disagree and dispute the result) to 0.9 (strongly agree and accept the result) through interactions with *MapView* and *TemporalView*.

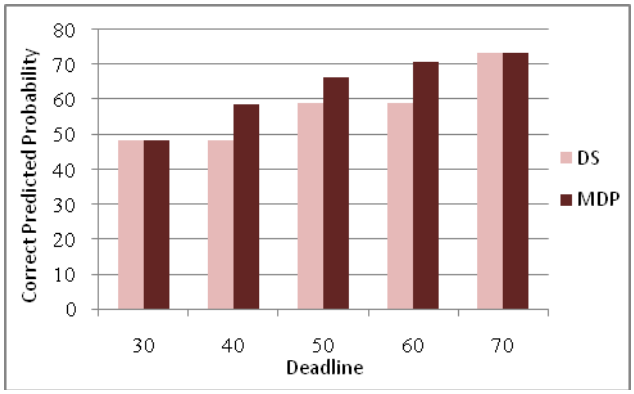

Figure 3. Comparison of correct predicted probability under different deadlines

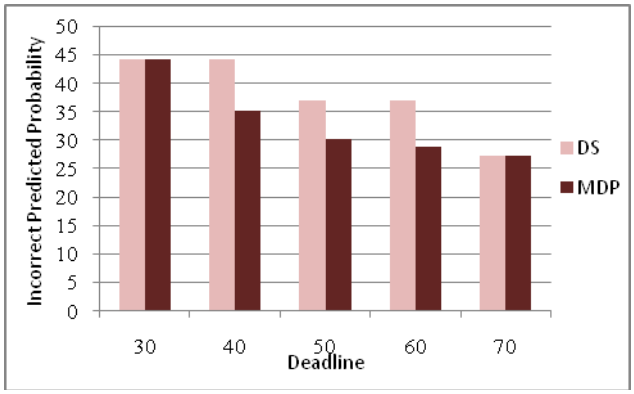

Figure 4. Comparison of incorrect predicted probability under different deadlines

We compare the predictive performance of the MDP policy and a Deterministic Schedule (DS) for task structures under different deadlines. DS is a deterministic process scheduler that builds a static schedule with the highest possible quality. The DS in this experiment is: {*HQ-Model-Option; Classification-Analysis; HQ-Map-View-Option; HQ-Map-View-Interaction-Option; HQ-Temporal-View-Option; HQ-Temporal-View-Interaction-Option*}.

Compared with a traditional DS, our MDP policy shows a significant improvement in assisting the users to predict the correct group name. Shown in both Figure 3 and Figure 4, we provide detailed comparisons on both cases with correct predictions and incorrect ones. Both charts clearly show the dynamic policy that the MDP provides allows users to get more correct probability results and fewer incorrect ones than if they use a DS. For instance, there was 10.31%, 7.07%, and 11.53% improvement in performance for the deadline of 40, 50, and 60 with the correct prediction respectively. The t-test values (0.000286, 0.010028, and 0.000695) are less than 0.05, which means that performance of MDP policy is statistically significantly different from the performance of DS. Therefore, the RESIN agent is able to assist analysts to make better responses especially on task deadlines 40 to 60 with MDP policy.

Noticeably, there is not much difference for deadline 30 and 70 since they are highly constrained and loosely constrained problem. Due to a tight deadline (30), the MDP solver cannot generate a policy better than the DS, while on a loose deadline (70), the DS has enough time to complete all methods, just as the MDP policy could. Overall, the MDP policy outperforms the DS throughout our entire task set.

4.2. Event Prediction with Time Series Analysis

We now describe experiments to predict occurrences of terrorist attacks within a particular following time period based on the historical events by employing time series analysis. Time series analysis [9] comprises methods that attempt to identify the nature of the phenomenon represented by the sequence of observations, or to predict future values of the time series variable based on known past events. Here, we employ the Box-Jenkins [10] procedure, a widely used and most efficient forecasting technique, especially for time series, for the analysis of terrorist events in GTD [5].We carry out the time series analysis for the prediction from two aspects: 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.

4.2.1. Entire GTD data

All terrorist events grouped by month of occurrence (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 based on this model. With the log transformation of the initial data variable, a unit root test [14], we obtained the series which satisfy prerequisites of the Box-Jenkins method that requires the time series to be stationary with constant mean value. From correlogram [10] and associated statistic values, the pattern of autocorrelation [11] can be captured by a model of auto regression (AR) with an order of 1, 2, or 3. With the comparison of the optional models based on three pivotal main parameters-Adjusted R-squared (the coefficient of determination), Akaike Info Criterion (AIC) [12] and Hannan-Quinn criterion [13]-we determined AR (3) to be the best model for the GTD data series.

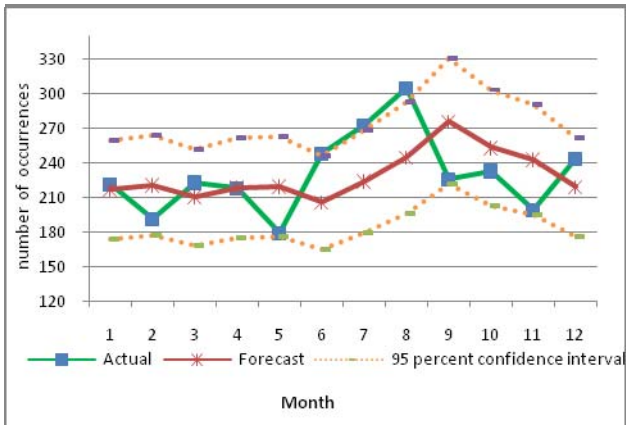

Figure 5. Actual and predicted number of terror occurrences per month from 1997.1 to 1997.12

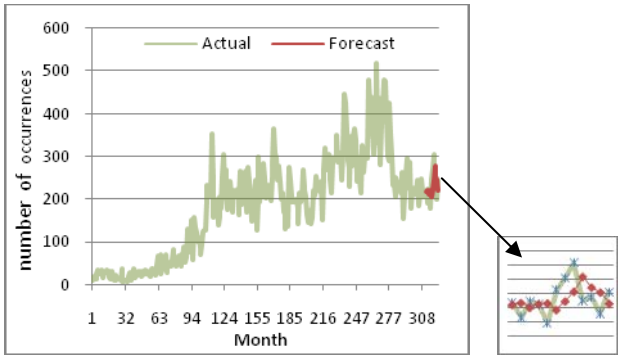

Figure 6. Actual occurrence from month 1th to 324th with comparison of predicted ones from month 313th to 324th

We noticed that the parameter of Mean Absolute Percent Error (MAPE) [11] for a static prediction is 15.51916 based on the model of AR (3). It is a significant index to measure the ability of forecasting for the model in statistics. This observation shows that the forecasting model we used here is good since the MAPE value is less than 20. Normally, if MAPE is less than 10, it is assumed that the model is highly accurate for the short time prediction [11].

Figure 5 indicates the close fit between the predicted amount of terrorist events and the actual occurrences. Taking a close look, the predicted value is almost the same as the actual one in April1997, although in August1997 there is a big disparity between the two. However, most of the 12 actual values are within the 95 percent confidence interval of the predicted values. Therefore, the auto regression model of order 3 is considered to be efficient in describing the occurrences of terrorist events in the GTD. Compared with overall data in GTD shown in Figure 6, our forecasting part is in close accordance with the actual curve which significantly proved that the approach we used here for the prediction with the overall GTD database is feasible.

4.2.2. Partial GTD data on Region & Group

The geographical circumscription is separated into six regions in the GTD and the worldwide attacks are carried out by 2404 terrorist groups respectively. Due to the missing data in the GTD database, we collected the occurrences of Region data by year while grouping the attacks of Group data by quarter for the purpose of prediction. We define Region data as terrorism occurrences categories by Region while Group data as terrorism occurrences categories by terrorist Group.

REGION	Mean	Std Dev	Predicted Time	Predicted Value	Actual Value
North America	-2.0000	31.38025	1997	17	31
Europe	20.34615	159.8849	1997	755.34615	514
Middle East/ North Africa	13.88462	227.5612	1997	194.88462	554

Table1. Occurrences of terrorist events for Region

GROUP	Mean	Std Dev	Predicted Time (Q1)	Predicted Value	Actual Value
Fatah	0.015873	1.689551	1992.Q1	1.015873	1
Hezbollah	0.050847	3.892627	1997.Q1	2.050847	4
ETA	0.058252	11.50944	1997.Q1	0	10
IRA	-0.010526	13.64403	1997.Q1	3.989474	7

Table2. Occurrences of terrorist events for Group

The stationary series for Region and Group were obtained by the transformation of one-order differencing [10, 11] to the initial data. For the correlogram analysis [12], these series are distributed randomly within the confidence interval, which indicates that they are purely

stochastic series that can not 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]. We use this method to perform the prediction on the Region and Group data.

Table 1 describes the results of the forecasting process on the Region data. There are only 27 years' data 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 quarterly data points for the Group, the observed standard deviations shown in Table 2 are much lower than the Region data shown in Table 1. Noticeably, there is not much disparity between the predicted values and actual values, for instance, the forecasting for the occurrences of Group Fatah in the first quarter of 1992 is almost the same as the actual value. This supports one initial hypothesis that GroupName (GNAME) is a good indicator for predicting terrorist event trends.

Thus, the approach applied here for random time series on the GTD is feasible and can provide a rough estimation for the possible value. However, the predicted value has a great dependence on the previous value, which means the method is limited within the short time forecasting.

5. Conclusion and Future Work

We have described a complex reasoning agent RESIN for predicting unknown or missing information in the GTD. We have equipped RESIN with the ability to predict future trends with time series analysis. Also, by integrating our agent with the visualization tool and the classification analysis tool, we have identified abstract representations of the tasks could largely improve the accuracy our automated system.

RESIN is a good start. However, there are still some interesting areas that we would like to investigate in the future. 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. Also, we attempt to enhance RESIN's prediction capacity especially for a long-range period by exploring other predictive modeling methods.

6. Acknowledgement

This work was sponsored by the Department of Homeland Security under the auspices of the Southeastern Regional Visualization and Analytics Center.

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* [Computer file]. ICPSR04586-v1, College Park, MD: University of Maryland [producer], 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2007-04-04.

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

[7] Quinlan, J. R., *C4.5: Programs for Machine learning*. Morgan Kaufman, 1993.

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

[9] Peter J. Brockwell, Richard A. Davis, *Time series theory and methods*, published by Springer, 1991

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

[11] Douglas C. Frechtling, *Forecasting Tourism demand methods and strategies*, Butterworth-Heinemann, 2001.

[12] Francis M.Mutua, *The use of the Akaike Information Criterion in the identification of an optimum flood frequency model*, Journal of Hydrological science, Vol.39, No.3, June 1994.

[13] Hannan, E.J. and Quinn, B.G., *The determination of the order of an auto regression*. Journal of the Royal Statistical Society B41, 1979, pp. 190–195.

[14] Nelson, C.R. and Plosser C.I., *Trends and random walks In Macroeconomic Time Series*, Journal of Monterey Economics, 10, 1982, pp.139-162.

[15] 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.