The Role of Blackboard-based Reasoning and Visual Analytics in RESIN's Predictive Analysis

Dingxiang Liu[1], Jia Yue[1], Xiaoyu Wang[2], 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
{dliu, jyue, xwang25, anraja, ribarsky }@uncc.edu

Abstract

Knowledge gathering and investigative tasks in open environments are very complex because the problem-solving context is constantly evolving, and the data may be incomplete, unreliable and/or conflicting. This paper significantly extends our previous work on a mixed-initiative agent by making it capable of assisting humans in foraging task analysis using AI blackboard-based reasoning, visualizations and a user interface. The agent's reasoning process involves leveraging the sequential execution of multiple knowledge sources, gathering large amounts of evidence, reasoning about incomplete and contradictory information, and supporting hypothesis tracking. The agent is equipped with the ability to adapt its processing to available resources, deadlines and their current problem-solving contexts.

1. Introduction

Due to the significant increase in collected data and the increased complexity of the reasoning process itself, performing investigative analytical tasks has become more challenging. These tasks typically involve identifying and tracking multiple hypotheses; gathering evidence to validate the correct hypotheses and eliminating the incorrect ones [17]. They also require the assistance from interactive visualizations, which enable analysts to explorer and preprocess large amounts of data. More importantly, the analysis tasks are often time critical and need to adopt appropriate approaches, which vary from straightforward methods to comprehensive investigations.

One critical task is to predict missing or unknown information about current events based on trends from the past. The prediction [11] could be influenced by the varying viewpoints of stakeholders and internal biases of the news stories and sources of data used for the analysis, which lead to the great uncertainty in the analysis domain.

To facilitate the task-solving process, we present an automated reasoning agent, RESIN which will determine predictions about a single event based on information from multiple and sometimes even conflicting viewpoints. RESIN stands for, a RESource bounded INformation gathering agent for visual analytics. RESIN builds on our previous work on TIBOR [8] and emphasizes on the blackboard reasoning and mixed-initiative reasoning aspects that will assist investigative analysts in performing viewpoint-based predictive analysis. RESIN leverages sequential decision making [1] and an AI blackboard system [3] to support hypothesis tracking and validation in a highly uncertain environment [5]. Also, adopting an interactive visual analytics tool, RESIN has the capability to pass information between analysts and itself, during the problem-solving process. In order to achieve this communication, RESIN is designed to be capable of handling three types of procedures: gathering of large scale, high dimensional data from varieties of sources, determining the type of processing to extract the data from these sources, determining appropriate visual analytical views for these data. Moreover, RESIN provides ways for the user to interact with its problem-solving process collaboration or even control it at every step, through a detailed user interface. By using RESIN, investigative analysts will now have access to automated support for their decision making and the capability for finding non-myopic alternate solution paths; and they will have a tool to study outlier data.

In this paper, we are using RESIN, applied on the Global Terrorism Database (GTD) [7], to perform such predictive analysis tasks. Using classification techniques, blackboard-based reasoning and the GTD Visualization Tool [13], we are able to make predictions based on existing historical data. We also provide confidence measurements for those predictions.

In the following sections, we describe the prediction problem and briefly introduce the blackboard system; then, we describe our blackboard-based decision-making process, followed by an experimental evaluation. Finally, we conclude the paper by summarizing our contributions to the community and enumerating open issues for future work.

2. The Prediction Problem

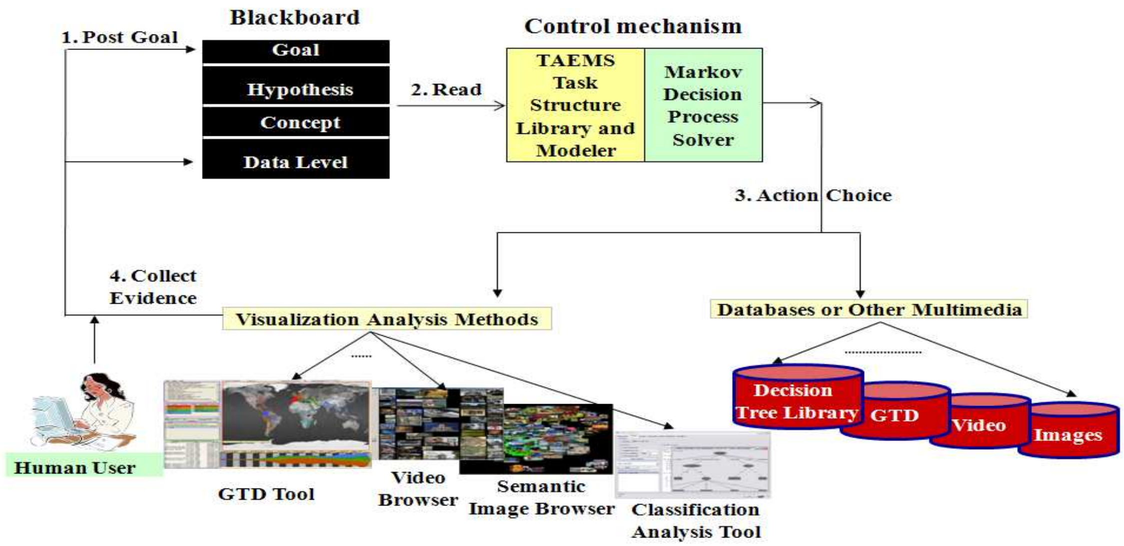

Figure 1: Functional Architecture for Predictive Analysis

RESIN consists of an AI Blackboard [3], a TÆMS [4] task structure library, a Markov Decision Process (MDP) [1] solver and heterogeneous knowledge sources, as shown in Figure 1. The AI Blackboard contains reasoning results from processing existing information, which includes raw data, various problem-solving states, partial solutions and current goals; TÆMS is an abstraction of the low-level execution model and captures uncertainty in outcome distributions, while the MDP is a probabilistic model, which captures the essence of sequential processes and is used to compute policies that identify, track, and plan to resolve confidence values associated with blackboard objects.

RESIN's Blackboard supports both opportunistic and planned verification of hypotheses and concepts. It is probable that, while gathering information to verify a concept supporting one hypothesis, the belief for a competing hypothesis increases. RESIN's Control Mechanism will model this possibility and then opportunistically redefine the problem-solving process to gather evidence to verify the competing hypothesis. RESIN agent's resource-bounded control mechanisms enable the agent to determine the appropriate databases and tools. To support reasoning about hypothesis, time, processor and other resource tradeoffs of various solution paths, RESIN has problem-solving models that are abstractions of the low-level execution model and capture uncertainty in outcome distributions.

The steps in Figure 1 illustrate the control flow of the predictive analysis process, which involves handling several issues: choosing the appropriate set of databases, analyzing the large scale, high dimensional data; generating the type of decision tree to extract and represent the data; determining appropriate interactive visualizations for these data; performing reasoning processes and generating final solutions.

The following is a description of a specific prediction process we applied in order to determine which terrorist group is likely to be responsible for a particular incident. 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 defined as an input vector with partial information about a single terrorist incident (Table 1). The task in Table 1 has six categories: TYPE, ENTITY, REGION, YEAR, NKILL, and WEAPON 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. Based on the blackboard's preprocessed result, the TÆMS task structure modeler generates an appropriate task structure and translates it to the MDP solver for action assessment (Step 2). Using dynamic programming, the MDP solver computes the optimal policy based on resources constraints (e.g. deadline) and generates the best action, which will trigger appropriate methods to perform predictive analysis (Step 3). 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 their results back to the AI blackboard (Step 4). Using these visualization results as well as previous analysis results, the blackboard will then propagate the evidence information and verify a specific hypothesis with an associated confidence value.

Table 1 Partial Terrorist Incident Description

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

We show that using the control flow described above, RESIN can largely enhance the accuracy of results in solving prediction problems; with the integration of visual analytics tools, RESIN will provide the capability for the user to manually perform tasks or even override agent's suggestions, interactively.

2.1. The Knowledge Sources

Knowledge sources (KSs) [3] are independent specialist computational modules that contain the domain knowledge needed to solve a problem. A knowledge source understands the state of the problem-solving process. At appropriate times, the knowledge source takes relevant information on the blackboard and makes contribution towards solving problem with its specialized knowledge. RESIN employs a set of knowledge sources, including Weka's C4.5 [10, 14], the GTD, and an investigative visual analytics system built on the GTD.

Weka [14] is an open-source data mining toolkit, developed at the University of Waikato in New Zealand, which implements a collection of classical machine learning algorithms (e.g. C4.5) and provides friendly graphical user interfaces for data preprocessing, classification and visualization. C4.5 is a classical machine learning algorithm introduced by Quinlan for inducing classification models from data. It is a successor to Quinlan's earlier ID3 algorithm [9] that introduces a number of extensions of the original ID3 algorithm, which include missing attribute values, continuous attributes, and pruning of decision trees and so on. RESIN is integrated with Weka to implement the C4.5 in order to automatically access and preprocess global terrorism data.

The Global Terrorism Database (GTD) [7] contains terrorist activities between 1970 and 1997, which have been collected by the Inter-University Consortium for Political and Social Research (ICPSR). This database contains approximately 60,000 records and each record uses 120 categories, such as responsible terrorist group, total number of persons killed and so on, to describe terrorist incidents.

The GTD tool [13] 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. Among all the views that this tool provides, there are two views that are significant in helping the reasoning process: MAP_VIEW and TEMPORAL_VIEW. MAP_VIEW provides straightforward geospatial information to depict terrorists' incidents while TEMPORAL_VIEW reveals their temporal trends and patterns, as well as the relative growth and decline among the patterns over time. The categorical information shown in Table 1 can be mapped into two views of the GTD tool with high interactivity, which plays an important role in the foraging analysis of our mixed-initiative agent. The world map is shown by the GTD tool interface. As the user zooms into the specific area, local detail will be presented in the map. We will discuss these knowledge sources in Section 3.

2.2. The Multi-level Blackboard Database

The AI blackboard [3] data structure is a global shared repository containing problems, elementary data, a set of partial solutions, contributed information, and other data, which is available to all KSs and serves as a communication medium. It is the kernel of the RESIN agent, providing a reasoning approach for the information that has been discovered and produced. 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. The layered hierarchy allows for explicit modeling of concurrent top-down and bottom-up processing, while maintaining a clear evidential path for supporting and contradictory information. The information at a given level is thus derived from the level(s) below it, and it in turn supports the hypothesis at higher levels.

2.3. The Control Mechanism

The control mechanism makes runtime decisions about the problem-solving process, specifically for a given hypothesis and resource (e.g. time) constraints, it will determine the databases and tools that need to be accessed. For RESIN, we have modeled this control mechanism using TÆMS-based uncertainty reasoning and a MDP-based [1] sequential decision problem. The essence of sequential decision problems is that decisions, which are made in resource-bounded, uncertain environments, can have both immediate and long term effects; the best action choice depends on the types of future situations. For prediction problem, the TÆMS task structure contains the various choices of visualization tools, classification algorithm, database, and levels of user interaction relevant to the particular query. Then the TÆMS task structure is translated into a Markov Decision Process solver by initializing a state set, identifying the possible actions to determine the optimal action choices, and expanding each possible outcome which is characterized by discrete quality, cost and duration values, as described in our previous paper [8].

For an automated agent to be accepted by the analyst community, this agent must provide analysts the ability to manually direct a search or override actions suggested by the control mechanism. The contingency plans built into the MDP policy will allow the control system to adjust dynamically to such overrides.

3. Implementation

In this section we present a detailed description of the RESIN agent's reasoning process on a prediction example, with the goal to determine unknown group name (GNAME) based on key tuples, as shown in Table 1. There are many categories for each incident in GTD. In order to obtain a reduced representation of the data set and maintain the integrity of the original data, we use a decision tree induction technique [16] to select some categories from GTD, which are most relevant categories to GNAME, and remove irrelevant or weakly relevant categories. To achieve the goal of this example, we selected 100 historical terrorism incidents as the training set.

Based on the input tuples, 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 tool. To justify the importance of user interaction in a mixed-initiative agent, the RESIN's task structure also provides some important user interaction options, such as *Map-View-Interaction-Option* and *Temporal-View-Interaction-Option*.

In the next step, the task structure is translated into a MDP solver by computing the optimal policy which in turn prescribes the most appropriate knowledge source to trigger. In this task, the policy triggers Weka to implement C4.5 [10] to generate the decision tree and predict the group name. Weka can not only generate the view of this decision tree based on algorithm C4.5, but also show the specific confidence value of the predicted value and optional value (if possible). This automatic processing of this knowledge source will provide a partial solution, which is posted onto blackboard (Step 4 in Figure 1). The information is then stored in the proper table in the RESIN's Blackboard described in Figure 2. The knowledge source C4.5 predicts that the group name may be Fatah with confidence value 0.75, along with one alternative solution that the group name may be Hezbollah, with confidence value of 0.25. Therefore based on its larger confidence value, Fatah is posted as C4.5's partial solution onto the blackboard.

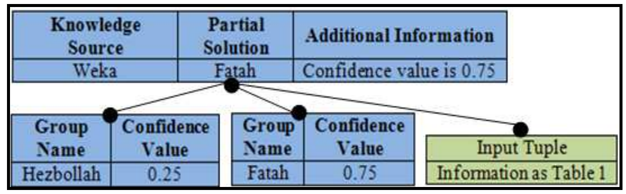

Figure 2: Partial Information on Blackboard (Knowledge Source is Weka)

Based on the MDP generated policy, RESIN will trigger another knowledge source, the GTD tool, to facilitate the Weka's results. When invoked, the GTD tool would first receive relevant information (Figure 2) from the blackboard to understand the current state 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 MAP_VIEW and TEMPORAL_VIEW. In Figure 3, the MAP_VIEW shows related incidents for those two Weka-predicted groups. The user can easily determine that even if these two groups' incidents share some similar geo-spatial distributions, they are quite different with respect to radius and ranges, suggesting unique terrorist activities. Also shown in Figures 4 and 5, the TEMPORAL_VIEW of those two groups revealed that although they have overlap during their active times, their main attack types are quite different. (The colors in the temporal views are keyed to different attack types, i.e., category TYPE.) The information from these two views will equip the human user with clues to estimate their confidence values and to adjust the predicted results.

With user interaction, the GTD tool posts those confidence values back to the blackboard and updates current information as shown in Figure 6 and Figure 7. The combination of each evidence value will be the contribution to final solution in the Goal level of the blackboard. The group name with highest confidence value will be posted onto the goal level of the blackboard. In this example, we define combination equation for confidence value is:

$$GC_{groupName} = \sum_{i=1}^{N} KS_i * KS\_Weight_i \quad (1)$$

Where N is the number of knowledge sources in RESIN.

In this example, we have three knowledge sources. So the equation is:

$$GC_{groupName} = DT * W_{DT} + MV * W_{MV} + TV * W_{TV} \quad (2)$$

Where DT is the confidence value from the data classification analysis (Weka); MV is the confidence value from the MAP_VIEW analysis; TV is the confidence value from the TEMPORAL_VIEW analysis; W_{DT} is the weight of data classification analysis (Weka); W_{MV} is the weight of MAP_VIEW analysis; W_{TV} is the weight of TEMPORAL_VIEW analysis. Through interactions with MAP_VIEW and TEMPORAL_VIEW, users could determine their confidence values based on the visual patterns shown in these two views. For example, as suggested by TEMPORAL_VIEW (Figure 5) Fatah has larger amount of assassination activities in 1992 compared to Hezbollah, user would agree with Weka's prediction by feeding back a high confidence value (0.7). Assuming W_{DT} = 0.6, W_{MV} = 0.2, and W_{TV} = 0.2, we obtain $GC_{Hezbollah}$ = 0.31 and GC_{Fatah} = 0.77 by applying equation (2).

Therefore, RESIN predicts the group name is Fatah with the confidence value 0.77. The multi-level blackboard database for the example is shown in Figure 8.

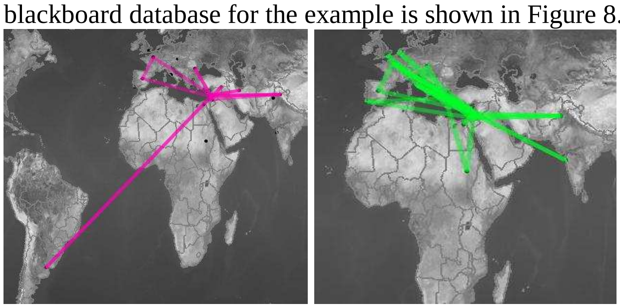

Figure 3: MAP_VIEW of Groups: 'Hezbollah' is on the left and 'Fatah' on the right

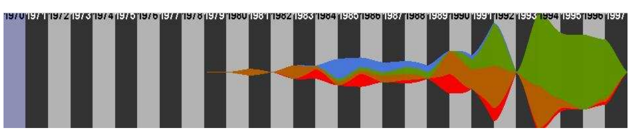

Figure 4: TEMPORAL_VIEW of Group 'Hezbollah'

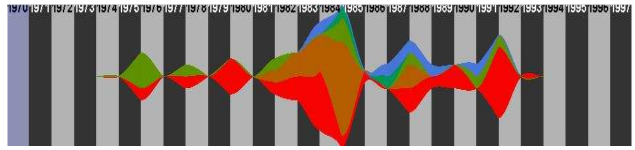

Figure 5: TEMPORAL_VIEW of Group 'Fatah'

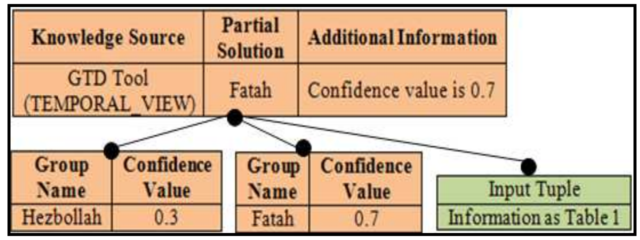

Figure 6: Partial Information on Blackboard (Knowledge Source is GTD Tool: TEMPORAL_VIEW)

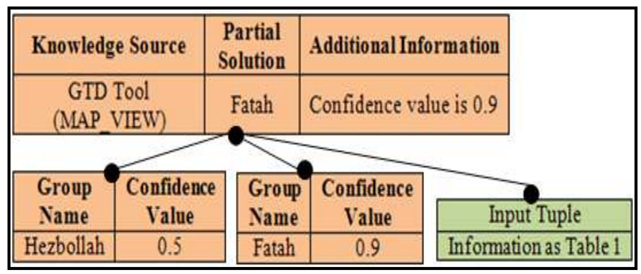

Figure 7: Partial Information on Blackboard (Knowledge Source is GTD Tool: MAP_VIEW)

4. Experiments

In this section, we describe two experiments to both assess the effectiveness of RESIN's blackboard-based reasoning mechanism, and also show the benefits from using the C4.5algorithm. These experiments are based on a training set of 2700 incidents selected from GTD and per each task we use the same ten incidents outside the training set, with different deadlines from 30 to 70 (ranging from a very tight to a loose deadline).

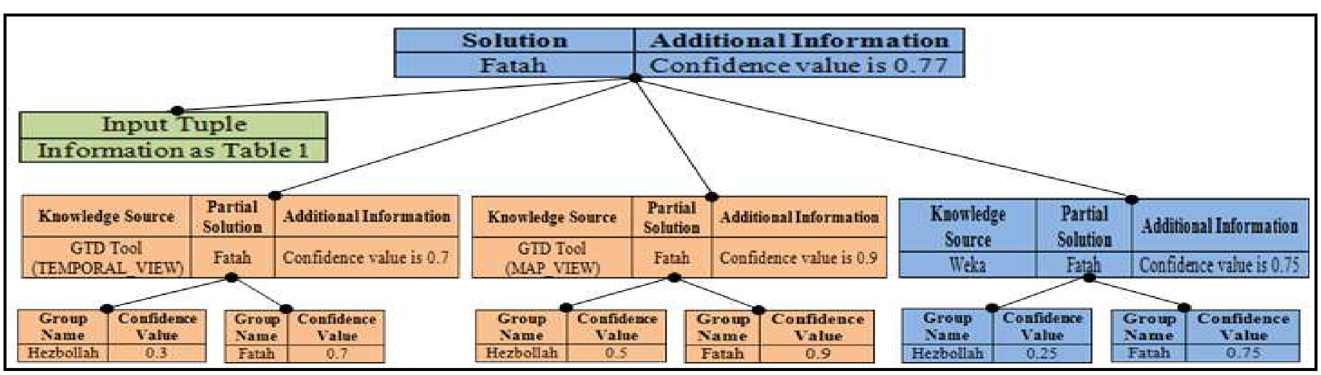

Figure 8: Multi-level Blackboard Database

First, we compare the predictive performance of the MDP policy and Deterministic Schedule for task structures under different deadlines. To make a fair comparison, we provide a static schedule with the highest possible performance, as similar task structure described in Section 3: {*HQ-Model-Option; Classification-Analysis; HQ-Map-View-Option; HQ-Map-View-Interaction-Option; HQ-Temporal-View-Option; HQ-Temporal-View-Interaction-Option*}.

By design, RESIN triggers Weka to employ C4.5 to generate the decision tree based on the training set and provide a partial solution for the test case. This partial solution is time critical and determines whether this task has enough time for the user to interact with visualization tools. If so, through interactions with MAP_VIEW and TEMPORAL_VIEW, users could determine confidence values towards initial predictions and post them back to the Blackboard, with values from -0.9(strongly disagree and dispute the result) to 0.9(strongly agree and accept the result). The Blackboard would then run the reasoning process again to generate final prediction results.

Compared with a traditional deterministic schedule, our MDP policy shows a significant improvement in assisting the users to predict the correct group name. Shown in both Figure 9 and Figure 10, we show detailed comparisons on both cases with correct predictions and incorrect ones. Both charts clearly show that based on the dynamic policy MDP provides, users could get higher correct probability results and lower incorrect ones than if they uses a Deterministic Schedule. These significant differences, especially on short task deadlines 40 to 60, indicate that the RESIN agent is able to assist analysts to make better fast responses within very limited time slots.

Noticeably, there is not much difference for deadline 30 and 70, due to their two extreme time factors. On an intense deadline (30), MDP solver cannot generate a policy better than the Deterministic Schedule, while on a loose deadline (70), the Deterministic Schedule gets enough time to complete all methods, just as MDP policy could. Overall, the MDP policy outperforms the Deterministic Scheduler throughout our entire task set.

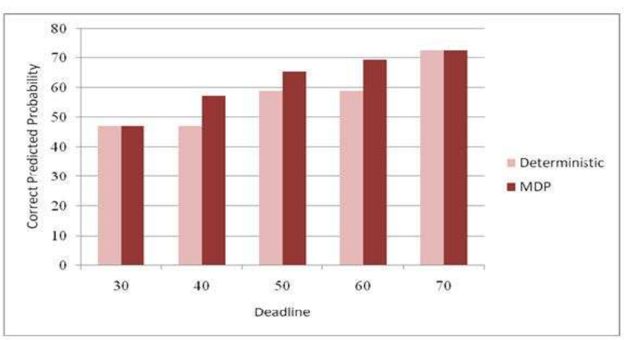

Figure 9: Comparison of correct predicted probability under different deadlines

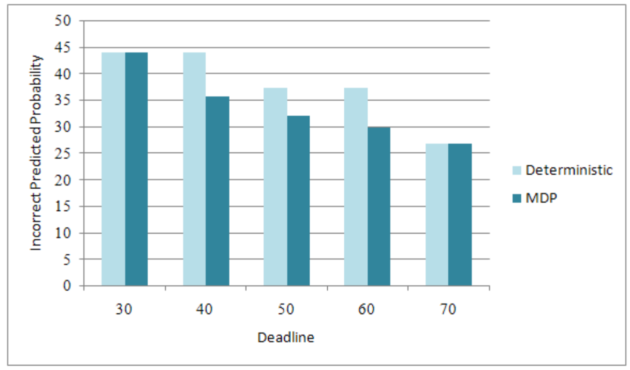

Figure 10: Comparison of incorrect predicted probability under different deadlines

Second, along with the comparison of predictions, we carried out another experiment to prove our design choice of applying the C4.5 algorithm to GTD, as shown in Figure 11. By using C4.5, RESIN gains 10% correctness on average, compared to three other widely used algorithms. Moreover, C4.5 still shows an uprising trend towards the end of this experiment, which indicates that this algorithm could also be used to deal with longer deadlines.

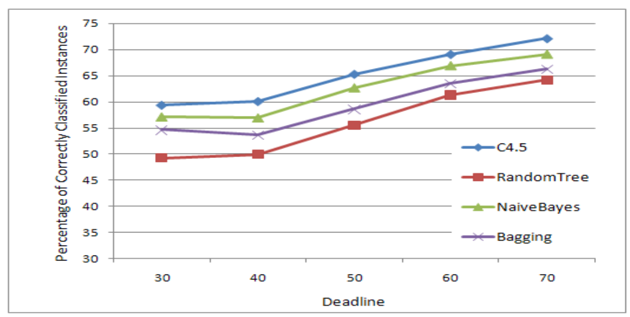

Figure 11: Comparison of various algorithms in Global Terrorism Database

5. Related Work

AI blackboard-based approaches use opportunistic reasoning in solving problems for which no deterministic solution strategies are known before hand, or for problems too vast for a complete exhaustive search. They also provide a design in which several specialized subsystems utilize their partial knowledge bases and strategies to build an approximated solution to the original problem. Hearsay II (1977) Speech Recognition [5] is the original blackboard which rejected a linear, data-driven analysis model in favor of a more complex, dynamic system. The blackboard functions as the global database of hypotheses that comprise the results of all inferences and predictions performed. Blackboard architectures also have been used as an underlying framework in which expert systems are embedded, for example BB1 [6], GBB [2] and the HAZOP [12] expert systems. Such structures have, however, been limited to handle traditional expert system reasoning, and they need to be generalized to handle other methodologies in the field of artificial intelligence, such as neural networks, fuzzy logic, and evolutionary computing. The blackboard in RESIN is an application in the AI domain which functions as a multi-level database for the information that has been discovered and produced thus far.

User interaction enables a model of an interactive process that can pass information between user and system. It is becoming increasingly popular for uses in a wide range of software applications especially some component-based architecture. In RESIN, there is a user interface that will enable analysts to collaborate with and, where necessary, to control the agent at every step of the problem solving process via a dashboard called the control panel which provides the current view as well future choices of the problem solving process.

6. Conclusion and Future Work

We have described a complex reasoning agent RESIN for predicting unknown or missing information in the Global Terrorism Database (GTD). We have identified abstract representations of the tasks to assist in the automated analysis as well as integrated the agent with the visualization tool, classification analysis tool and Global Terrorism Database.

RESIN is a good start but there are still some interesting areas on which we want to work in the future. First, we want to complete integration of the AI blackboard and a variety of knowledge sources as well as multimedia databases. Second, the prediction [11] could be influenced by the varying viewpoints of stakeholders and internal biases of the news stories and sources of data used for the analysis. The agent we will build is an automated reasoning agent that will facilitate an analyst's problem-solving process by determining predictions about an event from multiple and conflicting viewpoints. Also, we can use current time-series technologies [15] to analyze past events and predict future trends.

7. 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.

References

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

[2] D. Corkill, K.Q. Gallagher, and K.E. Murray, *GBB: a generic blackboard development system*, in: R. Englemore, T.Morgan (Eds.), Blackboard Systems, Addison Wesley,Reading, MA, 1988, pp. 503-518.

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

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

[5] Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy, *The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty*, Computing Surveys, Vol. 12, No. 2, June 1980.

[6] B. Hayes-Roth, and M. Hewitt, *BB1: an implementation of blackboard control architecture*, in: R. Englemore, T.Morgan (Eds.), Blackboard Systems, Addison Wesley,Reading, MA, 1988, pp. 297-313.

[7] LaFree, Gary, and Laura Dugan, *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.

[8] Dingxiang Liu, Anita Raja, and Jayasri Vaidyanath, *TIBOR: A Resource-bounded Information Foraging Agent for Visual Analytics*, Proceedings of 2007 IEEE/ WIC/ ACM International Conference on Intelligent Agent Technology (IAT 2007), pp 349-355, Silicon Valley, CA, Nov 2-5, 2007.

[9] J. Ross Quinlan. *Learning efficient classification procedures*, In R. S. Michalski, J. G. Carbonell, and T. M. Mitchell (eds.), Machine Learning: An Artificial Intelligence Approach. Palo Alto, CA: Tioga Press, 1983.

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

[11] Edna Reid, Jialun Qin, Wingyan Chung, Jennifer Jie Xu, Yilu Zhou, Robert P. Schumaker, Marc Sageman and Hsinchun Chen: *Terrorism Knowledge Discovery Project: A Knowledge Discovery Approach to Addressing the Threats of Terrorism*. ISI 2004: 125-145.

[12] R. Vaidhyanathan and V. Venkatasubramanian, *Experience with an expert system for automated HAZOP analysis*, in: European Symposium on Computer Aided Process Engineering, pp. 26-29, Rhodes, Greece. May 1996.

[13] Xiaoyu Wang, Erin Miller, Kathleen Smarick, William Ribarsky, Remco Chang, *Investigative Visual Analysis of Global Terrorism*, Journal of Computer Graphics Forum, Eurovis 2008.

[14] Witten, I. and Frank, E., *Data Mining: Practical Machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[15] Zhanggui Zeng, Hong Yan and Alan M. N. Fu, *Time-series prediction based on pattern classification*. AI in Engineering 15(1): 61-69, 2001.

[16] Jiawei Han, Micheline Kamher, *Data Mining : Concepts and Techniques*, Second Edition, Morgan Kanfmann, 2005.

[17] Ke-Bing Zhang, Mehmet A. Orgun, Kang Zhang and Yihao Zhang, *Hypothesis oriented cluster analysis in data mining by visualization.* Proceedings of the working conference on advanced visual interfaces, pp. 254-257, 2006.