

Distributed Coordination in Uncertain Multiagent Systems

Rajiv T. Maheswaran,
Information Sciences Institute
Univ. of Southern California
4676 Admiralty Way, #1001
Marina Del Rey, CA, USA

Craig M. Rogers,
Information Sciences Institute
Univ. of Southern California
4676 Admiralty Way, #1001
Marina Del Rey, CA, USA

Romeo Sanchez
Information Sciences Institute
Univ. of Southern California
4676 Admiralty Way, #1001
Marina Del Rey, CA, USA

ABSTRACT

We consider real-time multi-agent coordination in a dynamic and uncertain domain addressing both distributed state information and partial knowledge of the common reward function. The challenge is to find functional strategies when bounded rationality hinders the ability to encompass the values of possible sample paths of the system. This paper discusses a new approach based on assigning agents to monitor portions of the reward structure for which they aggregate and propagate appropriate profiles which compactly represent relevant information used for policy modification. This approach shows promise as an alternate and potentially superior technique with respect to current decision-theoretic and scheduling approaches.

1. INTRODUCTION

We address coordinated execution of activities of a multi-agent team in domains with uncertainty. Joint operations in military settings, large-scale disaster rescue, project/personnel management in global enterprise settings, and multiple-rover missions in science-discovery are some examples of dynamic execution environments where effective and efficient coordination is crucial to success. In these domains, we have uncertainty and sequential decision-making, partial state information, incomplete policy knowledge, subjective views of the team reward function, and environmental instability.

In this paper, we present a concept where agents are assigned responsibility for portions of the reward structure. This responsibility is to aggregate and disseminate information, stored in *profiles* in a neighborhood of reward nodes. The key is to create profile parameters that (a) can be computed and communicated quickly and (b) contain metrics that are immediately relevant to policy modification. This approach allows agents to collectively deliver effective approximations of the relevant global information to the appropriate agents in a timely manner. The benefits of this approach are verified by a comprehensive independent evaluation against extensions of currently prominent decision-

theoretic and scheduling schemes.

2. PROBLEM MODEL

Here, we present a model that contains all the challenging properties discussed earlier. The model is an instantiation and extension of the TAEMS framework [1]. Every agent in the team has a set of activities that it can perform but it can execute at most only one at a time. Each activity has probabilistic outcomes for duration and quality occur with probability and respectively: Each activity can be started only once and must begin after a release time and finish at or before a deadline in order to obtain positive quality. Only the agent that owns a method knows its statistics and status at all times. In addition, only the agent has current knowledge of its policy at all times. Thus, this model have incorporated uncertainty, decisions over time, partial state information and incomplete policy knowledge

The team reward is a function of the qualities of all activities, and the agents' objective is to maximize this reward at some terminal time. One way this function can be composed is with a tree where the activities are leaf nodes. Each non-leaf node is associated with an operator (such as *max*, *min*, or *sum*) which takes the qualities of its children as input. The output of the root node is the team reward function. Extensions include having a link between nodes where the quality of the source must be positive at the start time of all descendants of the target for the target to obtain positive quality. Additional types of links could modify the targets in other ways (disabling, facilitating, hindering). We consider the case where each agent sees all ancestral nodes of activities they own, and any nodes and links that connect to its activities and their ancestral nodes via directional links. Finally, environmental instability is introduced because during execution, the reward network, the temporal constraints and the distributions can change, so even an optimal policy for the original problem can be invalidated at run-time. Thus, this model incorporates all the complexities discussed earlier.

3. APPROACH

The key to solving this problem is to find a method to disseminate the relevant global information that agents need but do not have, while respecting bounded-rationality to produce real-time decisions (e.g., a naive centralization procedure would incur large computation cost and communication delays). The key to our solution is the creation of profiles for each node in the reward network. Profiles are a characterization of the substructure of the reward net-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AAMAS'07 May 14-18 2007, Honolulu, Hawaii, USA.
Copyright 2007 IFAAMAS.

work, rooted at the given node, that are compact. Instead of centralizing state information and reward structure for all activities in the underlying structure, which would be cumbersome and would not scale, profiles try to capture the critical information needed for decision-making with a minimal number of parameters. These parameters are calculated from input obtained only from the neighboring nodes, thus the interaction scales as a function of link density of the reward node.

3.1 Reward Node Assignment

The first step in our approach is to assign an agent to each node in the objective view. The agent will be responsible for updating and disseminating a *profile* for the node. Updates depend on the profiles on nodes that are neighbors of the node (i.e., connected to it via an ancestral or directional link). Currently, we assign nodes randomly among agents who can see thenode in their subjective view to avoid bottlenecks¹. The shaded nodes are those for which the assigned agent is responsible and the white nodes are those with which the assigned agent must communicate. The agent may be required to communicate with a node that is not in its initial subjective view. Assignment of reward node responsibility allows an agent to expand its subjective view to include all nodes one link removed from the assigned node.. The responsibility graph may also be disjoint. Whenever a profile is updated, it creates a flow of messages throughout the multi-agent system. The computation for the update and the type and targets of the messages are a function of which metrics within the profiles are being updated. The sources of information updates are status observations made by the leaf nodes (e.g. activity starts, ends, etc.) which update a local profile and begin the process of dissemination.

3.2 Profile Metrics

Now that question of *who* has been answered, we address the *what* with respect to profiles. Each profile contains a set of metrics. While the metrics differ in their meaning and usage, they have the following in common: (1) the metrics can be updated quickly when input from neighboring nodes arrive; (2) the metrics have significant and immediate use for policy modification.

The characteristics of this problem make it intractable to generate a policy that can prescribe optimal actions for all states. Thus, the team is given an initial schedule which is simply a mapping from times to actions without any instruction for contingencies. Thus, all contingency planning (i.e., policy generation and modification) must be in the system. As mentioned, it is too complex to plan for all contingencies (states). One must then employ a policy that is reasonable for some subset of contingencies. The evolution of the system will eventually move the team outside the subset where the policy performs well. The key is to discover the *when* and *how* to modify policies. The metrics in the profiles serve this purpose. We now discuss four particular profile metrics: schedule probability, potential probability, schedule importance, and potential importance.

Schedule Probability: The schedule probability metric is the likelihood that the current node will obtain positive quality under the current policies of all agents. This metric combines status information (whether and when a method

¹Alternate assignment strategies are topics under investigation

started executing), probability information (duration distribution), constraint information (release and deadline) and policy information (whether executing the method is part of a likely contingency). The significance of schedule probability is that it answers the *when* question. If the schedule probability of a particular node begins to fall, it means that we are moving into a regime where the current policies are deteriorating with respect to that node. The system can use this information to decide whether to begin a policy-modification procedure to rectify this situation.

Potential Probability: The potential probability metric is the likelihood that the current node can obtain positive quality under *any* policy set available to all the agents. This metric combines status information, probability information, constraint information but does not use current policy information. Potential probability profiles for non-leaf nodes are much more complex because their profiles depends on which subset of children one wishes to consider. Here, the critical criteria is determining when the potential probability of a task is zero or one.

Schedule Importance: Schedule importance reflects the marginal rate at which a node contributes to the overall team goal. i.e., if the schedule probability of success of a node increased a certain amount, by what factor of that amount would the schedule probability of success at the root increase, assuming all other schedule probabilities stayed constant. This method is very useful because performance depends heavily on agents finding the best possible way to allocate their execution time. When choosing between multiple local policy-modifications that involve adding an activity that was not part of the previous policy, it would be prudent to choose the one that contributes most to the team goal. The importance of an activity provides a value that allows for this discrimination among the available choices.

Potential Importance: The potential importance of a node characterizes if a node is *capable* of contributing to root schedule probability under *any* policy, which is a boolean condition. Intuitively, for a node to have potential importance, it must be able to contribute under some policy and either its parent or one of its targets must have potential importance. The utility of the potential importance metric is to determine when an activity can no longer contribute to the team goal. Whenever it cannot, we can modify either status (by aborting an executing activity that has no potential importance) or local policy (by removing the activity as a potential action in any contingency) to create opportunities for other activity insertions.

4. EXPERIMENTS

This research was part of substantial effort to address the problems with the all the complexities discussed in the introduction². It involved three main teams which included many universities, research laboratories and corporations both co-

²The work presented here is funded by the DARPA COORDINATORS Program under contract FA8750-05-C-0032. The U.S. Government is authorized to reproduce and distribute reports for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

operating and competing to design novel solutions.

Our Profile-Based Coordination (PBC) approach was compared to one that utilized a Distributed-Markov-Decision-Process (Dist-MDP) method and another, Flexible Interval Scheduling (FIS), which utilized temporal networks in a classical scheduling network. The Dist-MDP approach is based on calculating the appropriate subspace of the state space that the evolution of the system will follow and reasoning over that subspace. The rewards at the frontier of this space and policies when outside this space are obtained through greedy search [2]. The FIS approach utilizes a flexible interval for starting methods to absorb the uncertainty and uses incremental revisions to the schedule that attempt to maximize stability of agent activities [3]. While all three approaches are different in their solution methodology, they all attempt to solve the same problem, i.e., given a reward network and an initial schedule, create a multi-agent system that can adapt in real-time to uncertainty and dynamism, to maximize the quality achieved at the root of the reward network. All research groups submitted their respective agent systems to be evaluated independently on a large suite of test cases. There were over 2600 problems (reward networks) that were run multiple times for problems where the optimal solution was calculable. In addition, there was testing on larger-scale problems that had up to 10 agents and about 500 network nodes. The problems were not known ahead of time so no approximations or thresholds within the system could be tuned to the evaluation.

The results for the small-scale experiments are shown in Table 4. Each class denotes a particular type of experiment, e.g., “Synchronization” denotes experiments where timing was very important. The “Temporal Tightness” class were experiments where the temporal constraints (release and deadline) were very close to the largest durations of the activities, “Chains” denotes a large number of linked directional operators. These tests involved a handful of agents and a number of nodes sufficiently small such that a policy from a centralized MDP solver could be calculated. The numbers shown are the mean qualities of the various approaches as a fraction of the expected quality from the centralized MDP solution. The overall results, which were determined to be statistically significant, showed that our approach outperformed these extensions of classical solution methods.

Of more interest are the results for the large-scale experiments are shown in Table 4. In this evaluation, a score of 1 was given to the trial with the highest quality for a particular experiment. The scores of the other trial of all systems were scaled proportionally. Each system had a regime in which it performed well, however, the overall score of our system was significantly higher than the score of the other systems (.91 vs .77 and .62). We note how increasing scale has a drastic effect on the ranks of the performance of the systems. Most importantly, our system outscored the other systems by the largest difference (.99 vs .58 and .22) on the hardest problem category, “Hard Mix + Overlap” which combined many of the challenges of the other classes.

5. CONCLUSION

The keys concepts or profile-based coordination are the assignment of agents to nodes in a network representation of the team reward function and generating appropriate profile metrics that are computable and communicable quickly

Class	PBC	Dist-MDP	FIS
Dynamics	0.98	0.99	0.98
Interdependence	1.00	0.95	0.98
Chains	1.00	0.95	0.93
Temporal Tightness	0.98	0.94	0.90
Synchronization	0.99	0.92	0.81
New Task Arrival	0.86	0.96	0.95
Overall	0.97	0.95	0.91

Table 1: Results for small-scale experiments

Class	PBC	Dist-MDP	FIS
New Task Arrival	0.88	0.61	0.94
Window Tightness	0.84	0.93	0.64
Synchronization	0.97	0.87	0.70
Hard Mix	0.88	0.88	0.60
Hard Mix + Overlap	0.99	0.58	0.22
Overall	0.91	0.77	0.62

Table 2: Results for large-scale experiments

to create an information flow that has significant and immediate impact on policy modification. There are several directions for future research involving open questions in our approach including (i) optimal assignment of reward nodes to agents, (ii) better methods for evaluating the window effect, and (iii) automated determination for thresholds for which multi-agent policy modification is triggered. In addition, we developed our methodology on the premise that preventing failure (instances of zero quality at the root node of the reward network) was the key to maximizing quality. We are moving towards techniques that use root node quality maximization as the direct objective function. Despite the limitations in the aforementioned areas, our approach outperformed extensions of traditional approaches, especially in the most difficult regimes that were of most interest. We hope to develop profile-based coordination as a promising new direction for solving problems in this challenging setting of real-time multi-agent coordination under uncertainty and sequential decision-making, partial state information, incomplete policy knowledge, subjective views of the team reward function and environmental instability.

6. ADDITIONAL AUTHORS

Pedro Szekely, Information Sciences Institute, University of Southern California, 4676 Admiralty Way #1001, Marina Del Rey, CA, USA.

7. REFERENCES

- [1] K. Decker and V. Lesser. Quantitative modeling of complex computational task environments. In *AAAI*, pages 217–224, 1993.
- [2] D. Musliner, E. Durfee, R. Goldman, M. Boddy, J. Wu, and D. Dolgov. Coordinated plan management using multiagent MDPs. In *2006 AAAI Spring Symposium on Distributed Plan and Schedule Management*, 2006.
- [3] S. F. Smith, A. Gallagher, T. Zimmerman, L. Barbulescu, and Z. Rubinstein. Multi-agent management of joint schedules. In *2006 AAAI Spring Symposium on Distributed Plan and Schedule Management*, 2006.