AI Magazine Volume 15 Number 4 (1994) (© AAAI)

Comparative Analysis of AI Planning Systems

A Report on the AAAI Workshop

David E. Wilkins

■ The Workshop on Comparative Analysis of AI Planning Systems, held during the 1994 national AI conference, was lively and interesting. Both the theoretical and practical sides of the AI planning community were represented. Several papers contributed to the theoretical analysis of planning algorithms, and others showed the first steps toward convergence between such theoretical work and practical work on the system engineering aspects of working planners.

The Workshop on Comparative Analysis of AI Planning Systems, held during the 1994 national AI conference, was lively and interesting. Both the theoretical and practical sides of the AI planning community were represented, and both sides seemed to understand the other side better after the workshop. Several papers contributed further to the theoretical analysis of planning algorithms, either through frameworks for reconstructing planning algorithms or through empirical studies (Christer Backstrom, Linkoping University, Sweden; Subbarao Kambhampati, Arizona State University; Henry Kautz, AT&T Bell Labs; Craig Knoblock, University of Southern California/Information Sciences Institute [USC/ISI]; and Qiang Yang, University of Waterloo, Canada). Other papers showed the first steps toward convergence between such theoretical work and practical work on the system engineering aspects of working planners by developing generic frameworks to describe actual planners (Kambhampati; Andre Valente, University of Amsterdam, The Netherlands; and Yolanda Gil, USC/ISI).

Four general points of agreement emerged. First, the evaluation of planning systems is difficult. There are many different types and levels of evaluation, and it is necessary to be clear about what the goals of any evaluation are. Besides scientific evaluation within the AI community, there can be evaluation by end users of systems, government funding agencies, scientists in other disciplines, industry, and other entities. Each such evaluation is likely to focus on different criteria, and many criteria are subjective and qualitative. It was noted that comparing planners is similar in difficulty to comparing programming languages (in fact, the input specifications to a planner can be viewed as a programming language). Shlomo Zilberstein (University of Massachusetts) presented a number of evaluation measures.

Zilberstein called for a new approach to the evaluation of planners. An integrated system that executes or uses the generated plans should be evaluated instead of simply evaluating the plans that can be produced in isolation. For planning systems included in an agent situated in a changing environment, it's important to evaluate how the planner affects the performance of the agent in its environment. The planner is viewed as a source of information that is used by an execution architecture to select actions. A planner is only as good as the effect it has on the performance of the agent in its operational environment. This holistic view permits comparisons of agents that plan with those that do not (for example, agents using only reactive control with predefined procedures for situated activity). This view would also recognize the value of an exponential planning algorithm that can be combined with an executor and a replanner to produce an effective agent over a polynomial planning algorithm that cannot.

A second point of agreement was that hierarchical task network (HTN) planning, as done by systems such as SIPE-2 and O-PLAN, is more expressive and powerful than the precondition-achievement planning of algorithms and systems such as TWEAK, SNLP, and UC-POP. Adding hierarchical network expansion is not just an efficiency hack. Kambhampati presented theoretical results supporting this claim. Moreover, Mark Drummond (NASA Ames Research Center) claimed that the precondition-achievement planners that are the object of extensive study would never solve practical problems, and the best practical approach is to encode and use the planning knowledge of human experts, paying attention to the representation of the knowledge actually available from the experts. Although encoding expert knowledge is at the heart of HTN planning, there remains a considerable gap to bridge in using expert planning knowledge in our systems.

Third, it was generally acknowledged that common plan representations and ontologies, although difficult to design, would greatly aid comparison of planning systems. A common representation would allow formal comparisons among widely different planning technologies. One difficulty is avoiding a representation that is the union of all the representations in the different technologies. Papers by Gil and Valente were first steps in this direction.

Fourth, it is often impossible to extract an acceptable utility or evaluation function in practical planning problems. All the practical planning researchers attending the workshop concurred with this statement, based on their experiences across a variety of domains: military operations planning (Gil and David Wilkins, SRI International), manufacturing (Austin Tate, University of Edinburgh, United Kingdom, and Wilkins), machining (Caroline Hayes, University of Illinois), and trauma care (Bonnie Lynn Webber, University of Pennsylvania). Human evaluation is often subjective, with little thought given to formally specifying the evaluation criteria. (Operations research evaluation functions are simplified models of the actual situations and are not usually concerned with the many issues faced by planners, such as flexibility, robustness, and opportunities.) Although evaluation features can be defined, humans cannot seem to agree on a combining function. One obvious example is physicians and insurance companies that might disagree about whether it is worth performing a particular diagnostic test on a patient; similar disagreements can occur whenever evaluation is done by agents with different (or vague) value metrics. Several groups evaluated their plans by devising blind studies where panels of experts rated both human-generated and machine-generated plans. At least two conclusions can be drawn: (1) studying ways to extract evaluation knowledge from human experts would be worthwhile and (2) most AI search techniques assume an evaluation function (for example, A* needs an admissible heuristic function), thus making them hard or impossible to apply. The problem is worse than the typical problem of trying to find an admissible heuristic function for A* because it's hard to find any precisely definable mathematical expression of user preferences.

David E. Wilkins received his B.S. from Iowa State University in 1972, his M.Sc. from the University of Essex in 1973, and his Ph.D. from Stanford University in 1979; his thesis work centered on a chess program that used knowledge to replace and control search. Wilkins has since been at the SRI International AI Center, where he is currently a senior computer scientist. His research has centered on planning and reasoning about actions; knowledge representation; and design and implementation of AI systems, including SIPE-2.