

Empirical Methods in AI

Toby Walsh

■ In the last few years, we have witnessed a major growth in the use of empirical methods in AI. In part, this growth has arisen from the availability of fast networked computers that allow certain problems of a practical size to be tackled for the first time. There is also a growing realization that results obtained empirically are no less valuable than theoretical results. Experiments can, for example, offer solutions to problems that have defeated a theoretical attack and provide insights that are not possible from a purely theoretical analysis. I identify some of the emerging trends in this area by describing a recent workshop that brought together researchers using empirical methods as far apart as robotics and knowledge-based systems.

Twenty-five researchers gathered together during the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97) in Nagoya, Japan, for the Second Workshop on Empirical Artificial Intelligence. The workshop continued the work of a similar workshop held alongside ECAI-96 in Budapest, Hungary, the previous year. The workshop began with an invited talk by Henry Kautz (AT&T Labs). The rest of the workshop was arranged around two panels and two invited papers, designed to illustrate good empirical methods. Each panel had three panelists who kicked off a lively discussion by making some provocative remarks. The main goal of the workshop organizers was to limit the number of formal presentations and encourage discussion. The workshop format was highly successful at achieving this aim. I would recommend a similar format to other workshop organizers.

The debate during the workshop can broadly be divided into three categories: (1) past successes of empirical methods, (2) the design of computational experiments, and (3) the widespread use of random problems. The following summary necessarily offers just a partial description of the topics discussed during the workshop.¹

Success Stories

Empirical methods have been successful in recent years. Indeed, as Henry Kautz reminded the workshop participants, in the last year alone, the *New York Times* has reported two major empirical successes: (1) DEEP BLUE's defeat of Kasparov and (2) the computer-generated proof of an open problem in Robbins algebra. Pandurang Nayak (NASA Ames) described another highly publicized success, the diagnosis system for the Deep Space One spacecraft, which is based on a highly optimized satisfiability procedure. Although deciding satisfiability is intractable in general, this system generates plans in practice in essentially constant time for each step. It comes as quite a surprise to hear about real-time satisfiability testing.

Henry Kautz listed several reasons for the success of empirical methods. First, empirical studies are often an integral part of AI because systems can be too complex or messy for theory. Second, theory is often too crude to provide useful insight. For example, a problem might be exponential in the worst case but tractable in practice. Third, some questions are purely empirical. As Pedro Meseguer (IIIA, CSIC, Spain) pointed out during one of the panels, two search algorithms

might not be comparable theoretically because the nodes searched by one algorithm are not subsumed by the other, but empirical evidence might strongly suggest we prefer one over the other. Kautz identified several other reasons to use empirical methods. Experimental results might, for example, identify new computational phenomena. Much of the recent research in threshold phenomena (so-called *phase transitions*) has been empirical. The theoretical analysis of such behavior is currently far behind. Experiments can also suggest new theory and algorithms. For example, the large body of research into stochastic algorithms such as GSAT has been stimulated by empirical success. Theory is still a long way from explaining the success of local search on large satisfiability problems.

It should not be thought that theory and experiment are competing for success. Indeed, we should look for synergies between theoretical and empirical methods. Kautz illustrated such synergy with an example from description logics that addresses the trade-off between expressivity and tractability. The failure to be able to implement full equality efficiently led Schmidt-Schauss to prove its intractability.

Experimental Design

Jane Mulligan (University of British Columbia) offered some valuable advice to the workshop participants about experimental design in computational domains. She argued that experimental methods are an important tool in the analysis of complicated computational systems and can reveal behaviors that are not easily predicted. Factorial experiments indicate the significant factors, and resources can then be devoted to optimizing these factors. She suggested that the empirical analysis of system parameters and their effects on performance can provide us with robust solutions. Adele Howe (Colorado State University) illustrated how difficult it can be to compare algorithms designed for different goals with examples from her own research on web search engines. She asked if we run

One topic of great debate at this workshop was the widespread use of random problems in empirical studies.

the risk of bias because algorithm designers also usually perform the experiments and suggested that we should give people incentives to work on shared problems. High-profile competitions are one such incentive. Finally, she proposed the web as a good medium to make code and benchmarks available. Bernhard Nebel (Albert Ludwigs-Universität Freiburg) cataloged some of the problems that arise when comparing systems that are different in design. For example, central processing unit and memory use might be about the only way to compare graph planners and satisfiability planners. It is also not obvious how you compare systems that solve slightly different problems. For example, one planner might guarantee to return optimal plans, but another returns you any plan, and the third only returns plans of some fixed length. You also have to be careful about the choice of encoding because this choice can change the problem being solved. Finally, benchmark problems might have some feature such as symmetry that favors a particular approach.

Meseguer questioned how we should measure search cost. Because means alone are not satisfactory, how many statistics should we record? Are there implementation-independent measures we can collect (for example, the number of accesses to data structures)? Finally, he asked how we choose a benchmark. Should we separate soluble from insoluble problems? Do we just use problems from the middle of the phase transition? He suggested that benchmarks should include qualitatively different classes from the problem spectrum.

Random Problems

One topic of great debate at this (and, indeed, the previous) workshop was the widespread use of random problems in empirical studies. Research into phase-transition behavior has identified hard instances of random problems that are now routinely used to benchmark algorithms. Problems that are obviously overconstrained—or underconstrained—tend to be much easier to solve than critically constrained problems from around the phase transition in solubility. To counter Nayak's claim that he had never found a phase transition in real-life problems, Toby Walsh (University of Strathclyde) offered two examples: The first was an aircraft-scheduling problem for American Airlines, where the economics ensured that the problem was critically constrained in the number of aircraft. The second example was scheduling tracks at Gare de Lyon, where the problem proved to be so critically constrained that only a hardware interlock prevented a daily crash. John Slaney (Australian National University) and Sylvi Thiébaux (IRISA Rennes, France) then raised some interesting questions about phase-transition behavior in optimization problems ("what is the minimum-cost solution?") and its relationship to phase-transition behavior in the corresponding decision problems ("does a solution exist of cost less than some specified value?"). Using random blocks world planning instances as an example, they showed that the optimization problem might have a peak in hardness at a different point than the peak for the decision problem (typically, the peak in hardness for the decision problem is where about

50 percent or so of the problems have solutions) and offered some analysis of why this is so.

Jeremy Frank (NASA Ames) listed some of the pitfalls of using random problems. For example, some distributions give problems that are artificially easy, but random-number generators can be less random than we would like (especially when we are using them to generate just a single random bit). Meseguer also questioned if random models like that used by the constraint-satisfaction community are representative of real problems. Ulrich Hustadt (MPI, Saarbrücken) proposed that we should look for random problems that are *natural* (that is, they are similar to real problems met in practice), *hard* (that is, they are computationally expensive to solve), and *interesting* (that is, they should discriminate between our procedures). Unfortunately, it is usually hard to find all three properties.

Kautz suggested that we should add more structure into our random problems to make them more realistic. He offered quasigroup completion as a random problem with more structure. Walsh made a similar suggestion, proposing the use of *pseudoreal* problems. That is, we take real problems and perturb them randomly. We then have the benefit of large samples (as with random problems) and lots of structures (as with real problems). Of course, the question of how to choose base problems, and how to perturb them, is still left open. Finally, David Poole (University of British Columbia) had a more radical suggestion. He argued that the reason we can cope with the real world is its structure. Indeed, if it were completely random, we would not survive. AI should therefore be about exploiting structure in problems. He suggested that we need to find real problems in which structure can be exploited. Indeed, there might be much more structure than we would dare to put into a random experiment and claim that it is realistic.

Conclusions

The use of empirical methods in AI is flourishing at present. As the debate during this workshop illustrated, there

Computers & Thought



Edward A. Feigenbaum
& Julian Feldman

EDITORS

ISBN 0-262-56092-5 560 pp., index. \$18.00 softcover

The AAAI Press • Distributed by The MIT Press
Massachusetts Institute of Technology, 5 Cambridge Center, Cambridge, Massachusetts 02142
To order, call toll free: (800) 356-0343 or (617) 625-8569.
MasterCard and VISA accepted.

are many fine examples of results won through the use of empirical methods. Although experimentalists face many of the traditional problems found in other empirical sciences, there are several features of computational experiments that raise novel problems: For example, how do we measure performance in a machine- or implementation-independent way? The existence of a successful workshop series such as this one demonstrates that the field is facing up to such challenges. We should not rest on our laurels because good experimental practice is still being established in many areas. Nevertheless, the workshop allowed for a healthy dialogue between participants in diverse fields. It would, for example, be wrong to conclude that random problems will continue to be a major debating point. Indeed, it was agreed that the call for the next workshop would state that random problems is one topic that will not be discussed. I look forward to the next workshop, which will be held on 24 August 1998 as part of the Thirteenth European Conference on Artificial Intelligence (ECAI-98). Full details and an application form to attend the workshop are available from dream.dai.ed.ac.uk/group/tw/ecai98.html.

Note

1. See dream.dai.ed.ac.uk/group/tw/ijcai97.html for the call for participation.



Toby Walsh is a research fellow in the Department of Computer Science at the University of Strathclyde and an honorary fellow in the Department of Artificial Intelligence at Edinburgh University. He

received his B.A. from the University of Cambridge and his M.Sc. and Ph.D. from Edinburgh University. He has been a Marie-Curie postdoctoral fellow at INRIA (Nancy, France) and IRST (Trento, Italy) and an SERC postdoctoral fellow in the Department of Artificial Intelligence at Edinburgh. He is a founding member of the APES research group, a cross-university group of researchers dedicated to improving the use of empirical methods within AI (see www.cs.strath.ac.uk/Contrib/ipg/apes.html for more details). His e-mail address is tw@dai.ed.ac.uk.

New Proceedings from AAI Press

Proceedings of the Fourth International Conference on Artificial Intelligence Planning Systems

Edited by Reid Simmons, Manuela Veloso, and Stephen Smith

244 pp., 8-1/2 x 11 inches, index. \$40.00. ISBN 1-57735-052-9

The International Conference on Artificial Intelligence Planning Systems (AIPS) continues as the premier conference for showcasing new results in planning research. AIPS brings together AI researchers in all aspects of problems in planning, scheduling, planning and learning, and plan execution for dealing with complex problems. Papers in this proceedings range from new theoretical frameworks and algorithms for planning to practical implemented applications in a variety of domains.

Proceedings of the Eleventh International Florida Artificial Intelligence Research Symposium Conference

Edited by Diane Cook

504 pp., 8-1/2 x 11 inches, index. \$55.00. ISBN 1-57735-051-0

FLAIRS was founded in 1987 to promote and advance artificial intelligence research within the state of Florida, fostering interaction between researchers at colleges, universities, and industry. Since 1990, FLAIRS conferences have been broadened to include participants and papers from across North America and the world. This year's proceedings covers a wide range of algorithms, methodologies, and applications, featuring topics in uncertainty reasoning, natural language processing, machine learning, expert systems, knowledge representation, data mining, theorem proving, and AI education, with a variety of applications in areas such as accident prevention education, military planning, simulated race cars, and computer bridge.

Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology

Edited by Janice Glasgow, Tim Littlejohn, François Major Richard Lathrop, David Sankoff, and Christoph Sensen

234 pp., 8-1/2 x 11 inches, index. \$40.00. ISBN 1-57735-053-7

As with the previous ISMB conferences, this meeting provides a general forum for disseminating the latest developments in bioinformatics. ISMB is a multidisciplinary conference that brings together scientists from computer science, molecular biology, mathematics, and statistics. Its scope includes the development and application of advanced computational methods for biological problems. Relevant computational techniques include, but are not limited to machine learning, pattern recognition, knowledge representation, databases, combinatorics, stochastic modeling, string and graph algorithms, linguistic methods, robotics, constraint satisfaction, and parallel computation. Biological areas of interest include molecular structure, genomics, molecular sequence analysis, evolution and phylogenetics, metabolic pathways, regulatory networks, developmental control, and molecular biology generally. Emphasis is placed on the validation of methods using real data sets, practical applications in the biological sciences, and development of novel computational techniques.

To order, call 650-328-3123

or send e-mail to orders@aaai.org.

For further information visit our web site at

www.aaai.org/Press/

Published by The AAI Press

445 Burgess Drive

Menlo Park, CA 94025