The international planning competition series and empirical evaluation of AI planning systems

Derek Long and Maria Fox

University of Strathclyde, Glasgow, UK

Abstract. In this paper we consider the role of the International Planning Competition series in the evaluation of planners, both directly through the events themselves, and indirectly through the creation of resources and infrastructure. We also consider the problem of evaluation based on data collected both in the competitions and otherwise and examine some of the issues that arise in attempting to formulate and test hypotheses around the data.

1 Introduction

In 1998, Drew McDermott organised the first of what has become a biennial series of international planning competitions (IPC) [1,2,3,4,5]. Five planners competed in the first competition and many more have competed in each of the succeeding events. The competitions have stimulated a dramatic rise in performance of planners, with the achievements of planning systems having improved not only in terms of the speed to find plans, but also in terms of the range and complexity of the domain models they are able to manage. In particular, the 3rd IPC saw the development of temporal planning models [6], in the 4th IPC domain models were enhanced with derived predicates and the 5th (and most recent) IPC has seen the addition of soft constraints and trajectory constraints to the models of planning problems [7]. As a driver of research in planning, the IPC series has had impact beyond the immediate IPC events themselves, having encouraged the widespread use of a standard planning domain description language (PDDL), the creation of a large number of new benchmark domains and problem suites and a wide expectation that empirical evaluations of planning systems should now compare planners with the current collection of state-of-the-art systems as identified in the IPC series.

Competition series are increasingly common tools for stimulating research development and interest in a range of fields, including robotics (RoboCup and the Darpa Autonomous Vehicle Challenge), theorem proving, SAT solving and natural language processing. Competitions are extremely successful at focussing interest on specific problems and stimulating excitement in a field. A great benefit is that it is possible for a new approach to be applied by a completely unknown researcher in a field and, if it is successful, for it to rapidly become widely known and assimilated. They also have disadvantages: they can lead to loss of diversity as researchers seek to squeeze a little more performance out of whatever system won the preceding competition, they can cause research to focus on an artificial measure of performance which distracts from the real problems that the community might face and they can lead to stagnation. There are ways to mitigate these problems. In the planning community we have, so far, managed to achieve a rapid pace of development in the challenges set in the competition series, with a consistent pressure towards modelling realistic planning problems. There has also been a steady series of innovations in the field, although some techniques have become widely adopted standard approaches as a consequence of significant success in the competitions.

One of the challenges faced by the competition organisers is to define the basis on which winners are determined in each successive event. The changing challenges of the IPC series mean that evaluation has also evolved over the series. There is also an important difference between the choice of a winner in a competition and the more careful scientific evaluation of the comparative performance of different planning systems. In this short paper we briefly consider the problems and challenges that arise in using competition data as the basic of empirical evaluation of planners and we also discuss techniques which we believe to be useful in the comparative evaluation of such systems, based on our experiences of organising the 3rd IPC and performing the most extensive evaluation of the results that has so far followed any of the IPC series [3].

2 Evaluating Planning Systems

There are three dimensions that are most important in measuring the comparative performance of planning systems: the speed with which plans are produced, the coverage of the problems presented and the quality of the plans constructed in each problem that is solved. A planning problem is considered solved if a feasible plan is found. Optimality is not required (it is generally too hard to solve planning instances to optimality), but, nevertheless, the quality of plans is a relevant metric. By coverage we mean the proportion of problems solved and the balance of this proportion across different domains. For example, it is better to solve some problems from all domains than all problems from one domain but none from any of the others.

In the first competition, McDermott attempted to define, beforehand, a complex formula to attempt to balance these various factors and arrive at a winner. The formula was abandoned before the first round was complete and since then the evaluation, for the competition purposes, has been a rather more subjective one based on an intuitive balance between these factors.

The relative emphasis of the factors has changed a lot over the series. In the first two competitions, speed and coverage were primary, with quality being measured solely in terms of the number of steps in the plans and largely being placed behind the other factors. In the 3rd IPC a new extension of the competition language allowed problems to contain an explicit description of the metric that would be used to judge the quality of a plan, with the metric varying from problem to problem. This led to a far greater emphasis on plan quality and the three factors became more closely balanced. The 5th IPC has seen the emphasis on plan quality increase still further, with speed of planning dropping into a far less significant role. The argument has been that if a planner produces a plan in its time-limit (30 minutes in the 5th IPC) then the precise timing is not critical: programming tweaks, choice of data structures and programming language, compiler and so on can all have an impact on speed to the extent that when performance is measured in seconds the difference in speed is not a very reliable measure of the behaviour. Coverage is more important because it gives some insight into scaling behaviour and the power of a planner to handle a wide variety of language features and interactions between them.

The introduction of soft constraints made plan quality an even more important issue in plan production and a wide range of very important questions have been raised through the use of plan metrics and soft constraints in the evaluation of plans. For example, the use of soft constraints, or preferences, allows the specification of over-subscription problems, where not all of the specified goals can be achieved and the planner must determine which subset will achieve greatest reward at lowest cost.

Having identified the dimensions on which planners are evaluated, it is also helpful to observe that it has become standard practice in the planning literature to evaluate ideas by considering comparative performance. That is, a new technique is typically evaluated by empirical comparison of a planner sporting the technique with one that is not. In some cases, the base planner is the same for both data sets, with only the new enhancement differentiating the behaviours, while in other cases the new technique cannot be separated from the planner that uses it and then the comparison is between the new planner and some existing planning system, usually one recognised through the competition as a good performer.

It is worth emphasising, for the benefit of the reader who is unfamiliar with AI planning research, that the objective in much of the field is to construct a *general* planning system, capable of taking as input a declarative description of a domain and a problem instance for that domain (specifying the identity and initial configuration of objects in the domain and properties required as a goal) and producing a plan (an agenda of actions from the domain that will transform the initial state into one that satisfies the goal). The domain descriptions are problem-centred, rather than solution-centred, in the sense that they describe only what actions it is possible to perform and not the circumstances under which it might be *desirable* or *sensible* to perform them. Planning systems that rely on advice about when to apply particular actions are often called *knowledge-intensive* planners and have also been explored in the competition series, although less frequently.

We now turn to the critical question, which is how can comparison be drawn between two planning systems? The usual approach is to take a large sample of problems, from a wide range of benchmark domains, and run both planners on each problem. At this point, the hope is that one planner will show a sufficiently consistent improvement in all the useful dimensions that there is no need to do anything more than present the numbers in graphs and leave it to the reader to infer which system is better. In practice, there are many cases where things are not as convenient as this and the comparison in performance gives mixed messages, with sometimes one planner performing better and sometimes the other, sometimes with conflicting evidence when considering the different dimensions of performance. The consequence is often that claims about the performance of a new system must be qualified.

Several problems arise in drawing these comparisons. Firstly, although the benchmark problem set has increased since the IPC series began, nevertheless, empirical evaluations are restricted to a relatively limited set of domains. Some work has been carried out to analyse some of the properties of these benchmarks to determine important characteristics of them. For example, whether they are intrinsically hard problems [8] or yield good approximate solutions [9], whether they have natural structure that supports certain planning techniques [10] and whether they show particular kinds of symmetry [11,12].

In general, it is clear that good performance on the current set of benchmark domains is not a guarantee of good performance on new problems and planners can behave very differently when faced with a new domain. Secondly, the benchmark domains are associated with problem sets (typically anywhere between 20 and 50 problem instances are available for each domain) and these are designed to represent some sort of increasing challenge to planners. Unfortunately, it quickly becomes apparent that planners do not all agree on the relative hardness of individual problem instances and the intuitive notion of scaled problem difficulty does not always accord with the pattern of behaviour in the performance of planners. This problem is most acute in considering the behaviour across different planning domains: agreement within domains is stronger. Finally, the (necessary) use of a cut off time in evaluating the performance of a planner leads to an inevitable censoring effect in the data, where it is not always clear whether a planner might have returned a result if it had been given just a few seconds more to deliver it.

In practice, the empirical evaluation of planners rests on several assumptions: that the benchmark domains are somehow representative of a wider range of interesting problems, that the instances presented to the planners are representative of a wide range of problems in each domain, that the scaling behaviour can be usefully induced from a small set of problem instances and that the trends in relative behaviour of two planners can be induced from the sample on which they are based. Of course, several of these problems are typical of all empirical science and boil down to the robustness of the sampling strategy and the strength of the claims for generality made from the empirical evidence. The hypotheses that are typically of interest are assertions of consistent performance differences between planners in terms of one or more of the dimensions in which they are assessed. It is common to generalise from the samples to make claims about planner performance across “all planning problems”, if only implicitly.

2.1 Competitions and Scientific Evaluation

Our experiences in organising the 3rd IPC convinced us that it is very difficult to combine the generation of data in a competition context with the scientific evaluation of systems. A key difficulty is that the scientific method of constructing hypotheses and then carefully empirically testing them relies on freedom to respond to earlier results in order to construct new hypotheses and to have freedom to generate data from a wide sample space. In a competition, the key is to make an event that is informative and entertaining for the observers and that is practical in terms of time and technical demands for the competitors. The result of a competition is not peer-reviewed, but an entertaining judgement and the process depends on speedy assimilation of the data rather than slow and careful analysis.

After the 3rd IPC, we took the data that had been generated and performed several careful analyses, but we found that some of the questions we would have liked to answer could not be tackled with the data we had collected. Nevertheless, the competition generated more than 4000 data points (including about 4000 plans) between 13 different planners, across eight domains. Each plan included both a time value to produce it and a quality measure of the resulting plan. With this much data it is clear that there is scope to make some assessments about the relative behaviours of the various planners.

3 Empirical Methods and the 3rd IPC

We will now briefly review some of the techniques we applied in evaluating the results of the 3rd IPC and comment on the issues they raised. We believe that some of the techniques we applied in evaluating the data are very general and would be of interest and use in other efforts at empirical evaluation of AI systems.

3.1 Experimental Setup

The collection of the data was organised as follows: a single machine was used by all competitors. Competitors logged on to the machine remotely to run their planners on the problem sets. Domains were released with a limited time available to review them and test planners on sample problem instances (to check for any parsing bugs or other minor problems) before the main problem sets for the domain instance were released. Each domain had 20 associated problem instances of increasing size (measured in terms of the numbers of constants and related goals in each instance). The planners were allowed a maximum of 30 minutes on each problem instance. It was necessary to impose a limit because we wanted to collect a great deal of data in a short time.

3.2 Statistical Evaluation

Our approach was to construct a series of hypotheses and then to test them using familiar statistical techniques. For example, we formulated the null hypothesis that the data offered no basis on which to partially order the planners in terms of their performance, either in time or in quality. We used Wilcoxon rank sum matched pairs tests pairwise between planners, using a sufficiently small p value (we were more conservative than the Bonferroni correction and used p=0.001) to allow us to combine the results into a single tableau (partial order) between all the planners at a combined p value of 0.05.

An important advantage of the Wilcoxon test is that it is non-parametric, since it uses ranks rather than absolute data values. This matters a great deal, since the data points were collected for problems of (deliberately) increasing difficulty and the differences in performance between planners are affected by the sizes of the problems they are solving. In particular, on small problems the difference in performance is typically much smaller than on larger problems, as might be expected. Unfortunately, there is no way to normalise the differences according to problem difficulty because there is no reliable measure of problem difficulty — indeed, the question of how problem difficulty might be measured in order to correlate it with performance was one of the issues we considered in examining the results.

Since the performance differences are dependent on the problem difficulty as well as the planners, and the problem difficulties were not carefully controlled, the performance differences could not be expected to follow any particular distribution, forcing us to look at non-parametric tests.

Fully Automated	Strips	Numeric	HardNumeric	SimpleTime	Time	Complex
Depots	$F_{21,110} = 5.3$	$F_{21,44} = 5.48$		$F_{21,66} = 1.77$	$F_{20,63} = 2.14$	
DriverLog	$F_{19,100} = 17.1$	$F_{19,40} = 17.4$	$F_{19,40} = 4.05$	$F_{19,60} = 4.44$	$F_{19,60} = 4.63$	
ZenoTravel	$F_{19,100} = 21.7$	$F_{19,40} = 14$		$F_{19,60} = 9.4$	$F_{17,36} = 12.1$	
Rovers	$F_{19,80} = 4.54$	$F_{18,38} = 9.47$		$F_{19,60} = 4.25$	$F_{19,40} = 6.92$	
Satellite	$F_{19,100} = 7.36$	$\mathbf{F_{15,48} = 1.74}$	$F_{19,20} = 11.8$	$F_{19,60} = 3.6$	$F_{19,60} = 4.19$	$F_{19,60} = 3.78$
FreeCell	$F_{19,100} = 6.21$					
Settlers		$\mathbf{F_{5,6} = 1.6}$				

Fig. 1. F-values for the multiple judgments rank correlation tests.

The Wilcoxon test allowed us to check whether there was a consistent performance difference between pairs of planners, without considering its magnitude. A less powerful test, but one that is also non-parametric, is a simple proportion test in which we considered the proportion of problems in which one planner performed better than another.

Details of the outcome of our tests can be found in [3] — for the purposes of this paper, the important observation is that the the Wilcoxon rank sum matched pairs test was a valuable tool in examining the relative performance of systems on a data set with unknown distribution. Furthermore, this test is reasonably robust to the problem of cut off in the evaluation of relative timing performance because the ranks for these differences can be set to place them last in the series. The Wilcoxon test has been adopted by other authors in the planning field, influenced by our work, as a suitable test for performing pairwise performance comparisons between planners. We believe this to be a useful tool that might find wider application.

In order to check whether planners agreed on the relative difficulty of planning problems we used a "rank correlation in multiple judgments" test. We used the planners themselves as judges to determine how difficult individual problems were. In each test the n planners rank the k problem instances in order of time taken to solve. Unsolved problems create no difficulties as they are pushed to the top end of the ranking. The rank correlation tests for multiple judgements determines whether the independent rankings made by the n planners agree. The test statistic follows the F-distribution with $(k - 1, k(n - 1))$ degrees of freedom determining whether the critical value is exceeded. We formulated an explicit null and alternative hypothesis:

> **Null Hypothesis:** The planners differ in their judgements about which individual problem instances are hard within a given domain/level combination.

> **Alternative Hypothesis:** The planners demonstrate significant agreement about the relative difficulties of the problem instances within any given domain/level combination.

The results of the tests are shown in Figure 1 (repeated from [3]). The cells in the figure report the F values obtained (and the degrees of freedom used). In almost all cases the critical value was exceeded and the null hypothesis of non-agreement could be rejected for at least the 0.05 level. In a few cases (those reported in bold font) the critical value was not exceeded and no statistical evidence was therefore found of agreement between the planners about the difficulty of instances in the corresponding domain and level. It is interesting to note that the problematic cases are all within the NUMERIC level, for both fully-automated and hand-coded (knowledge-intensive) planners. This collection of problems includes actions with numeric effects and preconditions, rather than only logical conditions. In this competition they were a very new feature and performance in domains where they were used was quite variable.

Once we had established where there was agreement between planners about the difficulty of problems, we were also able to consider the relative scaling behaviours of planners on these problems (which is not possible if the planners do not agree on which problems are hard). In some cases the subset of problems on which there was consistent agreement about relative difficulty was too small to make comparisons of scaling behaviour reliable. This is one area where an opportunity

	FF	LPG		MIPS				Sapa	VHPOP	
		STRIPS	NUMERIC	STRIPS	NUMERIC	SIMPLE TIME	TIME	TIME	STRIPS	SIMPLE TIME
FF		0.36	⊙	0.87	0.93	—⊗—		⊗	0.93	⊗
LPG				⊙	0.52	0.51	0.61	0.58	0.44	0.48
MIPS								⊙	—⊙—	
Sapa	⊗			⊗	——⊙——				—⊙—	
VHPOP				⊙	⊗	⊙	⊗	⊗		

Fig. 2. Table showing correlation values, for fully-automated planners, between problem difficulty and difference in time performance, indicating scaling behaviour.

to examine the data before creating further tests might have led to more careful generation of test cases for scaling.

Our results in this case are shown in figure 2. The columns denote planners performing on individual types of domains (the competition had five or six variants of each domain, emphasising different language features in each case). The rows denote the planners and each cell contains the comparison between the corresponding pair of planners on a single domain variant type. We use ⊙ to indicate that there is insufficient agreement between the planners on the difficulty of domains, or the ranking of problems, for a comparison to be drawn. We use ⊗ to indicate that one of the planners in the pair being compared produced insufficient data for a comparison to be made. As can be seen, there are several places where the agreement between planners over the precise ranking of individual problems was too limited to support a further test for scaling performance (the situation denoted with ⊙ in figure 2). To avoid duplication of data we place entries as positive correlations only in the cell corresponding to the row for the planner favoured by the comparison. For example, FF is favoured in the comparisons with LPG, MIPS, Sapa and VHPOP. Once again, in these tests we used ranks rather than absolute values in order to avoid relying on any assumptions about the distributions. In this case, we used a Spearman rank correlation test. For those who are interested, a detailed discussion of the results and their interpretation is given in [3]. Our purpose here is to draw attention to the technique we used and the way in which it was applied.

4 Conclusion

In this paper we have discussed the IPC series which has proved highly influential in promoting the development of AI planning technology in the past eight years. We have briefly reviewed some of the issues that have arisen in the conflict between the goals of a competition and the goals of a scientific study. We have also described our own experiences in evaluating data generated in the 3rd IPC and we have discussed several of the statistical approaches we used in testing hypotheses about the relative performances of the competing planners. These techniques are, of course, widely known in the empirical sciences, but we believe that they should be of particular interest to the AI community, where they are, in our experience, less well known and even less widely used. Apart from stimulating huge growth in the power of planning systems, both in the expressiveness of the models they can handle and in the speed with which they can plan, the IPC series has also been central in promoting a more rigorous approach to empirical evaluation of planning systems and the use of more extensive data collection and evaluation. This is a healthy trend that should be encouraged in planning and in the wider AI community.

References

1. McDermott, D.: The 1998 AI planning systems competition. AI Magazine **21** (2000)

2. Bacchus, F.: The 2nd International Planning Competition home page. http://www.cs.toronto.edu/aips2000/ (2000)
3. Long, D., Fox, M.: An overview and analysis of the results of the 3rd International Planning Competition. Journal of AI Research **20** (2003)
4. Hoffmann, J., Edelkamp, S.: The Deterministic Part of IPC-4: An Overview. Journal of AI Research (JAIR) **24** (2005) 519–579
5. Gerevini, A.: The 5th international planning competion. ICAPS'06 Report (2006)
6. Fox, M., Long, D.: PDDL2.1: An Extension to PDDL for Expressing Temporal Planning Domains. Journal of AI Research **20** (2003)
7. Gerevini, A., Long, D.: Plan constraints and preferences on PDDL. In: ICAPS Workshop on Soft Constraints and Preferences in Planning. (2006)
8. Helmert, M.: Complexity results for standard benchmark domains in planning. Artificial Intelligence **143** (2003) 219–262
9. Helmert, M., Mattmueller, M., Roeger, G.: Approximation properties of planning benchmarks. In: Proc. of the 17th European Conference on Artificial Intelligence (ECAI 2006). (2006)
10. Hoffmann, J.: Where ignoring delete lists works: Local search topology in planning benchmarks. Journal of Artificial Intelligence Research **24** (2005) 685–758
11. Hoffmann, J., Gomes, C., Selman, B.: Structure and problem hardness: Goal asymmetry and DPLL proofs in sat-based planning. In: Proc. of ICAPS'06. (2006)
12. Fox, M., Long, D., Porteous, J.: Abstaction-based action ordering in planning. In: Proc. of International Joint Conference on AI (IJCAI). (2005)