Planning Domain Modelling Competition

Simon Dold

University of Basel, Switzerland
simon.dold@unibas.ch

Abstract

The international planning competition (IPC) is a recurring event that compares the performance of planners and awards the best ones. The evaluation is based on a set of benchmark problems that describe interesting planning problems in PDDL (planning domain definition language).

There is a general interest in high-quality benchmarks. Not only for the IPC but for the research of automated planning in general. We explore the possibility of a planning domain modeling competition that produces them. In this work, we discuss the desirable properties of benchmarks and how they could be evaluated.

Introduction

The first international planning competition (IPC) was introduced in 1998 to qualitatively compare different planning systems and measure the progress in the field (McDermott 2000). Additional goals of the competition were exerting pressure on the planning community and providing benchmark sets for evaluation. McDermott (2000) suggested to make the design of domains and generators a part of the IPC. Since an IPC can not be held without domains to evaluate the planners on, we want to explore the idea further how we could put this design into a competition.

Each new IPC issues a call for domain submissions. Recently, there has been additional motivation with the introduction of the Outstanding Domain Submission Award, recognizing exceptional contributions like Organic Synthesis (Matloob and Soutchanski 2016) as the award winner of IPC2018 and Quantum Circuit Layout Synthesis (Shaik and van de Pol 2023) as the award winner of IPC2023.

Both are submissions that model problems from different research areas. The IPC could gain more relevance outside of the planning research field if it solves interesting problems for other research domains (such as the ones receiving the outstanding domain submission award). If someone produces a PDDL domain for their problem, they would be motivated to polish it and send it to the new competition to (i) score high in the competition for providing an interesting problem that someone cares about (for that reason they produced a PDDL implementation in the first place) and (ii) guide the planners to become better in that domain as it could be used as a benchmark to evaluate planners.

This guidance is archived by tradition that, all benchmarks used for evaluation are public after the event and used by participants to test their planner for the next iteration of the IPC and for experiments in planning related papers.

An additional motivation to submit planning domains exists for the participants of the IPC. They can submit domains that work very well on their planner submission (or very badly on others) to improve their chances of winning the IPC. In the SAT Competition, this is handled similarly, with the difference that participants are *required* to submit benchmarks. Such submissions allow the participants to influence the final set of benchmarks and steer the competition in a favorable direction for them. This was also emphasized in the IPC call for domains as a motivation to provide a high-quality domain submission to be actually chosen as an IPC benchmark.

In this work, we lay out the qualities a good planning domain submission should have and mention the difficulties of rating them. Afterwards, we propose a high-level process to find the best submissions and discuss the option of using the winners of the planning domain modeling competition (PDMC) as benchmark domains for the IPC.

PDMC Submissions

PDMC submissions should be domains that could be used for the IPC. Vallati and Vaquero (2015) provided a list of desirable properties of a selection protocol for the IPC benchmarks and the benchmarks instances themselves. First, we look at the part about the benchmark instances. These are properties we would also like for PDMC submissions.

- **Challenging.** The problems are not too hard for any planner to solve nor trivially easy.
- **Interesting.** They describe problems of real-world situations.
- **Diverse.** A variety of different kinds of problems are described.

We want to look at these additional properties which are also desirable:

- **Natural encoding.** The objects and action/axioms reflect the problem with little to no auxiliary mechanism.

- **Adjustable.** The submission includes a generator that allows one to tweak important parts and control the size and difficulty of a problem in fine granularity.
- **Intrinsic Difficulty.** The difficulty should arise from the structure of the problem, it is not a simple problem scaled up.
- **Tricky.** They provoke an unintuitive difficulty or a common shortcoming of planners.
- **Optimality Bounds.** There is a domain-specific solver to provide optimal plans. Alternatively, there is a formal argument to provide a lower/upper bound to the optimal plan cost.

The SAT competition is similar to the ICP as it compares solvers for boolean satisfiability (SAT) problems. The IPC allows participants to submit domains where the SAT competition requires each participant (team) to submit benchmark instances. It also demands that (some of) the submissions are in an interval of hardness. Hard enough for a baseline solver to require more than one minute but not too hard such that the submitted solver would timeout in the competition time. The PDMC could also demand the submissions to be in such an interval. This partially outsources the generation of **challenging** instances from the IPC organizers to the PDMC participants.

Not all instances should be so challenging that only the top-performing solver can tackle them. It is desirable to have instances of different difficulties. They provide meaningful scoring even for domains that are not the strength of a solver. One natural way to create tasks in a variance of difficulties is to use an **adjustable** generator. The important part of being adjustable is that key aspects are expressed as parameters. As an example, we look at the sliding tile domain. Considering a generator of square-shaped sliding tiles problems would not be as adjustable as one with parameters for the two dimensions of rectangle-shaped sliding tiles problems. The former has a too coarse granularity providing little room between the too-easy and too-hard problems. The latter allows to generate many problems of different interesting difficulties.

There is no reason to expect that two teams submit the benchmark instances that model the same kind of problem. Which provides the organizers with a **diverse** collection of benchmark instances. The diversity here lies in the kinds of problems that are modeled. There could be a bias to problems that are easy for certain kinds of approaches but reveal bottlenecks for others. For example, if most participants would use planners that work on a lifted representation, then we could expect many benchmark submissions of domains that are hard to ground as this would provoke other representations to get out of memory. This could be done for other representations, algorithms, or heuristics as well.

For that reason IPC organizers should also consider a diversity in the planning approaches that correspond to the benchmark submission.

This can in part be prevented by including domains with an **intrinsic difficulty** for the benchmarks. Such a filter would prevent domains similar to Childsnack with 1000 ingredients as its difficulty arises from the size, not from the domain itself.

This is in part opposed to **tricky** domains. For example, the domain Gripper from the first IPC has no intrinsic difficulty but provokes planners to explore many equivalent paths. Such domains are also valuable as they motivate further research to deal with these issues.

However, with only these properties the IPC could evolve in a degenerative direction, only focusing on artificial benchmarks that focus on different bottlenecks of planning systems. This would decouple the IPC from the actual goal of developing a planner: providing a tool to solve someone's problems. For that reason, benchmarks should also contain **interesting** benchmarks in the sense of having a real-world interest. Using benchmarks that are interesting for other research areas in the IPC provides synergy between them and the planning research area as we mentioned above. (i) It allows these areas to steer the development of planners in a direction beneficial for them, as the problems of previous IPCs are commonly used to evaluate further developments in planning. (ii) Planning systems gain more relevance in other research areas.

With the outstanding domain submission from IPC2023 and IPC2018, this is additionally fertilized.

Another desirable property of a benchmark submission is that the domain is **naturally encoded** in the sense that few auxiliary actions or predicates or other "tricks" are used.

This pulls in an opposite direction than the ICKEPS2016 (Chrpa et al. 2016) competition, where it is one of the goals to have an encoding that is beneficial for planners. This steers the result to encodings with less accidental complexity (Haslum 2007) which could be a less natural way to describe the problem.

One could ask the question: Who is responsible to deal with the accidental complexity, the person developing the solver or the person modeling the problem? From the perspective of the ICKEPS2016 competition, it is the person modeling the problem. In this work, we hold the person developing the planner responsible. In practice, both have to put in the effort.

Additionally, submissions have to be manageable by planners. For that it should be constrained what language features are allowed. Otherwise, one submission could use PDDL features that are not supported by most planners or even introduce their own features. This could be prevented by a whitelist of language features that will be extended over the years.

So far we did not consider any particular track of the IPC but looked at it in general. However, for the optimal track in classical planning, an additional property is desired for a benchmark submission. That is access to **optimality bounds** because without them it is hard to score the performance of an optimal planner. Using the best solution found by a competitor puts the IPC into an awkward situation that breaks the independence of irrelevant alternatives. With knowledge of the optimal solution length, this issue can be avoided.

Rating Scheme

Most of the mentioned properties require human grading. This is a key difference between the PDMC to the IPC. We

want to propose a grading scheme to make the planning domain modeling competition more transparent. In the end, there will always be some subjective bias. This is unavoidable by the nature of human grading.

One submission can explain their **intrinsic difficulties** by an argument showing the problem belongs to a certain complexity class e.g. NP-hard. In this metric, we would put for example the Termes domain above the Childsnack domain.

It is hard to show how **interesting** a domain is in an unbiased way. One way is to provide references to others who tried to solve/optimize the modeled problem. Another is to give examples of where these solutions would be beneficial.

Similarly, for the **natural encoding** we can not simply quantify this. However, we can look for specific qualities. For the natural encoding that would be: Are there auxiliary actions/variables? Do the parameters of the action schema match the human intuition of what is needed for this action? Are the goal condition and the preconditions convoluted?

To show that a submission is in fact **challenging** and **adjustable** an experimental evaluation of at least one of the IPC planners from the previous instance could be used. A finite set of concrete instances should be dominated by a generator. Including a domain-specific solver for that domain in the experimental evaluation could show that the domain is **tricky** if the general planners struggle with it but a planner that "knows the trick" solves it easily.

For the **optimality bounds** it is obvious that a tighter bound dominates. Formal arguments about upper and lower bounds are better than no bounds at all. A domain-specific solver providing an optimal plan would dominate a function that simply returns the optimal solution cost. However, both would need a (formal) argument as to why the solution is indeed optimal.

The property **diverse** is analogous to novelty in paper submissions. A variant of an already established domain should score lower than a completely new one.

Process of Rating

In the end, some humans have to review the submissions. This could either be a committee (the organizers of the PDMC or a separate group), the peers, or crowdvoting.

Crowdvoting could end in unwanted criteria being the main deciding factor (Wilson, Robson, and Botha 2017). We do not expect this to end in degenerative voting behavior as it is a rather small community where the individuals have an incentive to get a good result from the voting. This was also proven with the best demo award of previous ICAPS (International Conference on Automated Planning and Scheduling) instances, which was also decided by a vote of the community. However, in this work, we do not further explore this direction. Nevertheless, we want to point out that this would implicitly add a new property to the list. Namely the presentability of the domain.

A process that seems more fitting is peer reviewing. Using peer reviews would make the PDMC a new ICAPS track. The peer review could (analogous to the main ICAPS track) be used to sort out weak submissions by rejecting them and accepting the better ones. Additionally, the reviewers can nominate someone for the best submission award. The final decision would be done analogous to the ICAPS main track.

The committee would be similar to the meta-reviewers in the peer-review solution. But they would have deeper insights into all submissions. However, the downside is that they would not be able to submit a domain themselves. This makes it hard to find volunteers for that role as they should be interested in domain modeling but do not want to submit to the PDMC themselves. Therefore a peer review process seems to be the best solution for the PDMC.

PDMC Winners as IPC Benchmark Domains

With the winners of the PDMC, we have a set of domains that can be used for the selection protocol for the IPC benchmarks (Vallati and Vaquero 2015). The desirable properties of the selection protocol are:

- **Transparency.** The process is understandable and reproducible by others.
- **Generality.** It is independent of the set of submitted planners or domains.
- **Unbiased.** It does not provide an advantage for some systems.
- **History-aware.** The process should disincentive overfitting the planners to benchmarks of previous IPC instances.
- **Progress-driven.** It provides an incentive to improve the submitted planners or submit new ones.

Do we get these by using the winners of the domain modeling challenge as the benchmark domains?

By stating the properties on which a domain submission is evaluated the **transparency** can be ensured. Furthermore, an open peer-reviewing tool can be used to increase transparency even more.

As the process is not dependent on a specific planning system or domain it is **general**. Once the details are settled it can be reused for further iterations of the competition.

The process is not fully **unbiased** as some planning systems are favored over others. A planning system with similar strengths and weaknesses as most other planning systems have an advantage against one that is very different from the others. With a large number of planning systems with strength X and weakness Y, it is more likely to have many submissions that focus on X but not on Y. A planning system, where Y is the strength but X is the weakness, has more submissions that are disadvantages. This could crystalize the development of planning systems in one direction. For that reason, the winners of the PDMC should not be the entire set of benchmark domains unless they are sufficiently diverse.

We expect many new domain submissions that are different from the benchmark domains of previous IPC instances, which contributes to **history-awareness**.

With multiple submissions that focus on different strengths and weaknesses, the development of each planning system is steered to work on its weaknesses and advance in its strengths. Additionally, a growing whitelist of

language features enforces further development to support them. Therefore, using winners from the PDMC keeps the IPC **progress-driven**.

The hope is to receive submissions similar to the workshop paper Academic Advising Planning Domain (Guerin et al. 2012). They provide explanations for many of the desirable properties we listed above and describe a generator for instances of different sizes and difficulties.

Conclusion

We discussed the potential synergy of a planning domain competition with the international planning competition, the planning research area, and other research areas. Additionally, we provided a high-level description of a process to rate the submissions and important dimensions for the overall evaluations.

The goal of this work is to spark a discussion within the planning community. We want to hear the concerns, suggestions, and comments from different voices and form the PDMC according to the interests of potential participants, reviewers, and organizers.

Acknowledgments

This research was supported by TAILOR, a project funded by the EU Horizon 2020 research and innovation programme (grant agreement no. 952215) and by the Swiss National Science Foundation (SNSF) as part of the project "Lifted and Generalized Representations for Classical Planning" (LGR).

References

Chrpa, L.; McCluskey, L.; Vallati, M.; and Vaquero, T. 2016. ICKEPS 2016 The Fifth International Competition on Knowledge Engineering for Planning and Scheduling. Web site: https://ickeps2016.wordpress.com/.

Guerin, J. T.; Hanna, J. P.; Ferland, L.; Mattei, N.; and Goldsmith, J. 2012. The Academic Advising Planning Domain. In *ICAPS Workshop on the International Planning Competition*, 1–5.

Haslum, P. 2007. Reducing Accidental Complexity in Planning Problems. In *Proc. IJCAI 2007*, 1898–1903.

Matloob, R.; and Soutchanski, M. 2016. Exploring Organic Synthesis with State-of-the-Art Planning Techniques. In *ICAPS 2016 Scheduling and Planning Applications woRKshop*, 52–61.

McDermott, D. 2000. The 1998 AI Planning Systems Competition. *AI Magazine*, 21(2): 35–55.

Shaik, I.; and van de Pol, J. 2023. Optimal layout synthesis for quantum circuits as classical planning. In *Proceedings of International Conference on Computer Aided Design (ICCAD)*, 1–9. IEEE/ACM.

Vallati, M.; and Vaquero, T. 2015. Towards a Protocol for Benchmark Selection in IPC. In *ICAPS Workshop on the International Planning Competition*, 34–38.

Wilson, M.; Robson, K.; and Botha, E. 2017. Crowdsourcing in a time of empowered stakeholders: Lessons from crowdsourcing campaigns. *Business Horizons*, 60(2): 247–253.