Hands-on planning: Evaluating planners

ICAPS 2013 Summer School. Perugia, Italy

Sergio Jiménez Celorrio   Carlos Linares López

Planning and Learning Group
Universidad Carlos III de Madrid

June, 7, 2013

Evaluating planners

Which planner should I buy?

Evaluating planners

Planner	Total
lama-2011	216.33
fdss-1	202.08
fdss-2	196.00
fd-autotune-1	185.09
roamer	181.47
fd-autotune-2	178.15
forkuniform	177.91
probe	177.14
arvand	165.07
lama-2008	163.33
lamar	159.20
randward	141.43
brt	116.01
dae-yahsp	101.83
cbp2	98.34
yahsp2	94.97
yahsp2-mt	94.14
cbp	85.43
lprpgp	67.07
madagascar-p	65.93
popf2	59.88
madagascar	51.98
cpt4	47.85
satplanlm-c	29.96
sharaabi	20.52
acoplan	19.33
acoplan2	19.09

Table: Final scores sequential satisficing track IPC-2011.

Evaluating planners

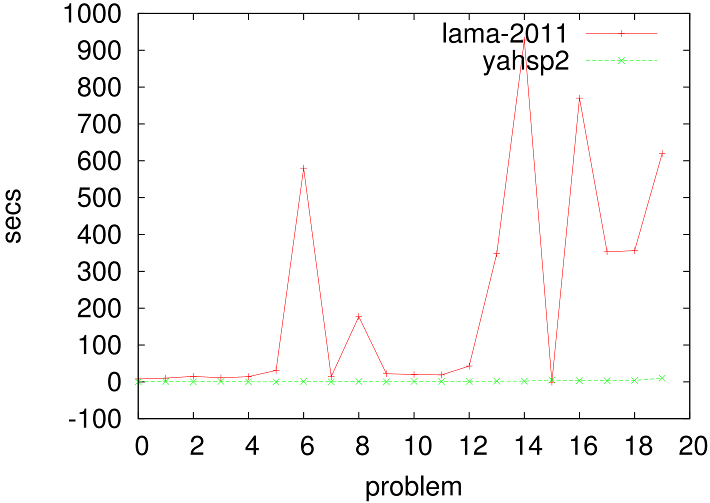


Figure: Time first solution *Transport* domain sequential satisficing track IPC-2011.

Planning task

Which planner should I buy?

- Which planning task do I need to solve?
- Under which conditions?

Outline

1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. Statistical Tests
5. Evaluation reports
6. Homework

Planning task

Which planner should I buy?

- which planning task do I need to solve?
  - how do states, actions and plans look like?

Planning task

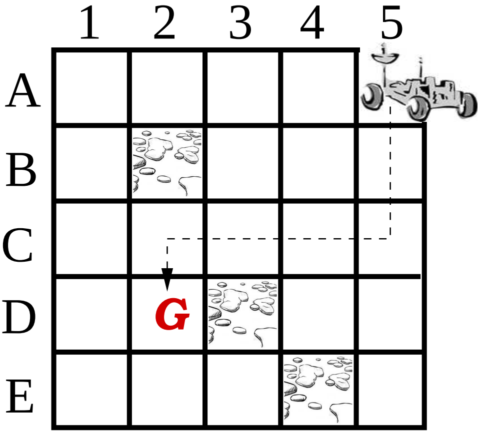


Figure: Reaching **D2** starting from **A5** with actions $\rightarrow, \leftarrow, \uparrow, \downarrow$.

Planning task

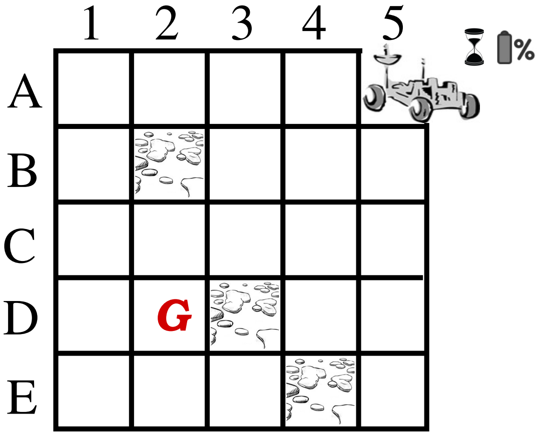


Figure: Reaching **D2** starting from **A5** with actions $\rightarrow, \leftarrow, \uparrow, \downarrow$.

Planning task

Planning task

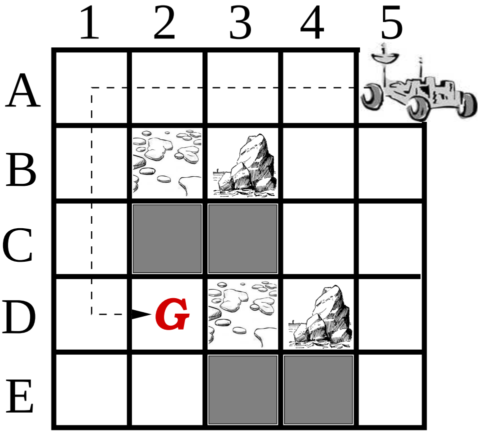


Figure: Partially observable states.

Planning task

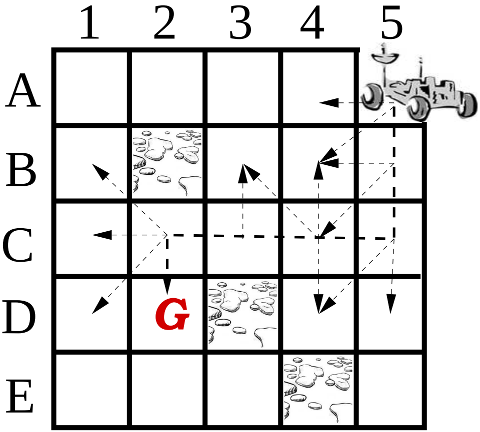


Figure: Non-deterministic actions.

Planning task

Planning task

Different planning task according to different

- States and Actions
  - Resources, time, uncertainty
- Plans
  - Satisfaction and optimization requirements

Planning task

Which planner should I buy?

- which planning task do I need to solve?
    - Planning model
    - Performance metric
    - Benchmark

Planning task

Planning task

Planning performance metrics quantify the achievement of scientific/engineering requirements.

Planning task

Different metrics used in planning (they are not exclusive)

- IPC metrics,
  - Number of solved problems
  - Time first solution plan
  - Plan length or plan make-span
  - Plan quality, IPC-2008 and IPC-2011 [Linares et al., 2013]
- other planning metrics,
  - flexibility [Nguyen and Kambhampati, 2001]
  - stability [Fox et al., 2006]
  - diversity [Nguyen et al., 2012b]
- other desired planning requirements,
  - justified actions [Haslum, 2012]
  - agents decoupling [Brafman and Domshlak, 2013]
  - ...

Evaluating planners

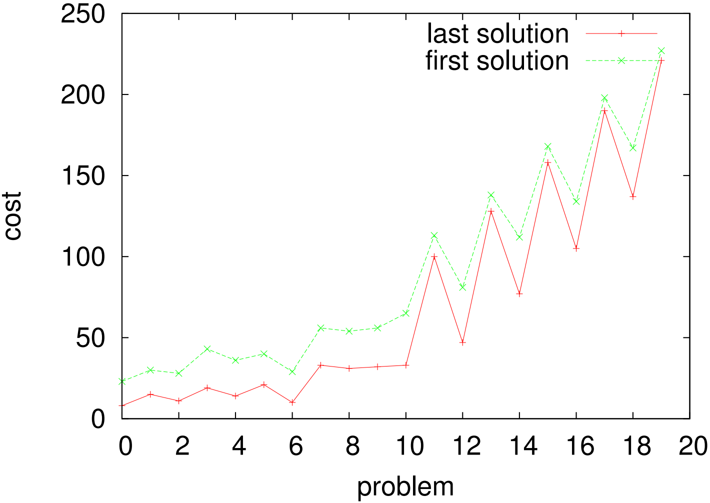


Figure: Cost of the first and last solutions found by lama-2011, *Openstacks* domain sequential satisficing track IPC-2011.

Evaluating planners

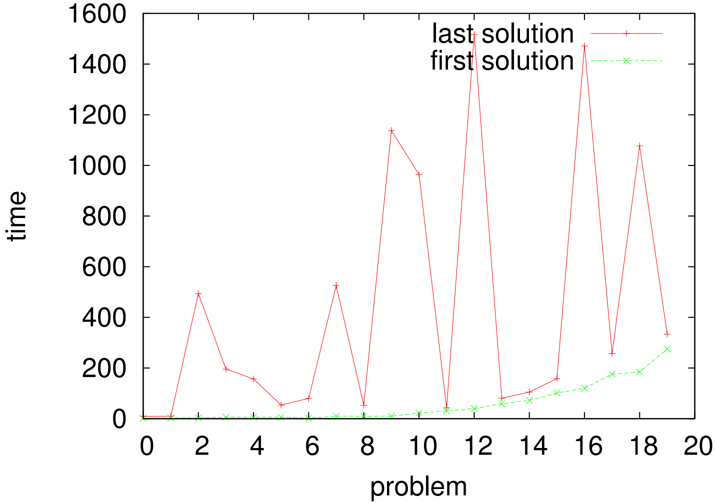


Figure: Time of the first and last solutions found by lama-2011, *Openstacks* domain sequential satisficing track IPC-2011.

Planning task

Which planner should I buy?

- which planning task do I need to solve?
  - Planning model
  - Performance metric
  - Benchmark

Planning task

Benchmarks that verify the achievement of the scientific/engineering requirements

- Overall performance
- Stress tests, specific challenges

Planning task

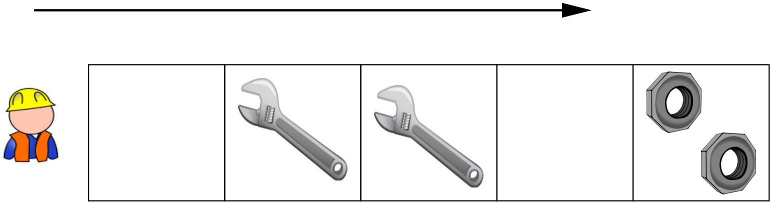


Figure: The *spanner* domain from the learning part of the IPC-2011. The worker must pick up the wrenches and tight the nuts. Challenging for planners based on the 'delete lists' relaxation since wrenches get broken after one use and the worker cannot comeback.

Summary

Which planner should I buy?

- which planning task do I need to solve?
    - Planning model
    - Performance metric
    - Benchmark

Outline

1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. Statistical Tests
5. Evaluation reports
6. Homework

Evaluation setup

Which planner should I buy?

- Which planning task do I need to solve?
- Under which setup? [Howe and Dahlman, 2002]

Evaluation setup

Evaluation setup

- Score function
- Computational resources
- Domains/problems

Evaluation setup

Different metrics used in planning (they are not exclusive)

- IPC metrics
  - Number of solved problems
  - Time first solution plan
  - Plan length or plan make-span
  - Plan quality, IPC-2008 and IPC-2011 [Linares et al., 2013]
- other planning metrics,
  - flexibility [Nguyen and Kambhampati, 2001]
  - stability [Fox et al., 2006]
  - diversity [Nguyen et al., 2012b]
- and more desired requirements,
  - justified actions [Haslum, 2012]
  - agent decoupling [Brafman and Domshlak, 2013]
  - ...

Evaluation setup

Planner	Total
lama-2011	216.33
fdss-1	202.08
fdss-2	196.00
fd-autotune-1	185.09
roamer	181.47
fd-autotune-2	178.15
forkuniform	177.91
probe	177.14
arvand	165.07
lama-2008	163.33
lamar	159.20
randward	141.43
brt	116.01
dae-yahsp	101.83
cbp2	98.34
yahsp2	94.97
yahsp2-mt	94.14
cbp	85.43
lprpgp	67.07
madagascar-p	65.93
popf2	59.88
madagascar	51.98
cpt4	47.85
satplanlm-c	29.96
sharaabi	20.52
acoplan	19.33
acoplan2	19.09

Planner	Total
lama-2011	250.00
probe	233.00
fdss-2	233.00
fdss-1	232.00
fd-autotune-1	223.00
roamer	213.00
forkuniform	207.00
lamar	195.00
fd-autotune-2	193.00
arvand	190.00
lama-2008	188.00
randward	184.00
brt	157.00
yahsp2	138.00
yahsp2-mt	137.00
cbp2	135.00
cbp	123.00
dae-yahsp	120.00
lprpgp	118.00
madagascar-p	88.00
popf2	81.00
madagascar	67.00
cpt4	52.00
sharaabi	33.00
satplanlm-c	32.00
acoplan2	20.00
acoplan	20.00

Table: **Quality** and **Coverage** rankings of the sequential satisficing track IPC-2011.

Evaluation setup

Planner	Total
lama-2011	216.33
fdss-1	202.08
fdss-2	196.00
fd-autotune-1	185.09
roamer	181.47
fd-autotune-2	178.15
forkuniform	177.91
probe	177.14
arvand	165.07
lama-2008	163.33
lamar	159.20
randward	141.43
brt	116.01
dae-yahsp	101.83
cbp2	98.34
yahsp2	94.97
yahsp2-mt	94.14
cbp	85.43
lprpgp	67.07
madagascar-p	65.93
popf2	59.88
madagascar	51.98
cpt4	47.85
satplanlm-c	29.96
sharaabi	20.52
acoplan	19.33
acoplan2	19.09

Planner	Total
lama-2011	155.21
probe	154.63
fdss-2	137.22
fd-autotune-1	129.51
roamer	118.81
lamar	115.54
forkuniform	113.62
fd-autotune-2	103.79
randward	102.06
yahsp2-mt	101.96
lama-2008	101.66
fdss-1	99.57
yahsp2	99.40
madagascar-p	77.71
arvand	77.39
brt	74.31
lprpgp	72.62
cbp2	59.92
cbp	56.84
daeyahsp	48.73
madagascar	48.52
popf2	41.93
cpt4	32.41
satplanlm-c	16.58
sharaabi	13.91
acoplan	9.05
acoplan2	8.12

Table: **Quality** and **Time** rankings of the sequential satisficing track IPC-2011.

Evaluation setup

Quality score for satisficing planners (IPC-2008 and 2011)

- $Q(planner, problem) = \frac{BestCost(problem)}{BestCost(planner, problem)}$
- $Q(planner) = \sum_i Q(planner, i)$
- $BestCost(problem)$ must be the optimal on the contrary the ranking computed with this score can be altered

Evaluation setup

	PlannerA	PlannerB	Optimal
000	10	20	10
001	20	40	5
002	100	60	60
003	110	80	80
mean	60	50	
median	60	50	

Table: Quality of best plans found for problems 000-003.

without optimal solutions $\mathbf{Q}(\mathbf{PlannerA}) > \mathbf{Q}(\mathbf{PlannerB})$

$$Q(\text{PlannerA}) = \left(\frac{10}{10}\right) + \left(\frac{20}{20}\right) + \left(\frac{60}{100}\right) + \left(\frac{80}{110}\right) = 3.327$$

$$Q(\text{PlannerB}) = \left(\frac{10}{20}\right) + \left(\frac{20}{40}\right) + \left(\frac{60}{60}\right) + \left(\frac{80}{80}\right) = 3$$

with optimal solutions $\mathbf{Q}(\mathbf{PlannerA}) < \mathbf{Q}(\mathbf{PlannerB})$

$$Q(\text{PlannerA}) = \left(\frac{10}{10}\right) + \left(\frac{5}{20}\right) + \left(\frac{60}{100}\right) + \left(\frac{80}{110}\right) = 2.577$$

$$Q(\text{PlannerB}) = \left(\frac{10}{20}\right) + \left(\frac{5}{40}\right) + \left(\frac{60}{60}\right) + \left(\frac{80}{80}\right) = 2.625$$

Evaluation setup

Evaluation setup

- Score function
- Computational resources
- Domains/problems

Evaluation setup

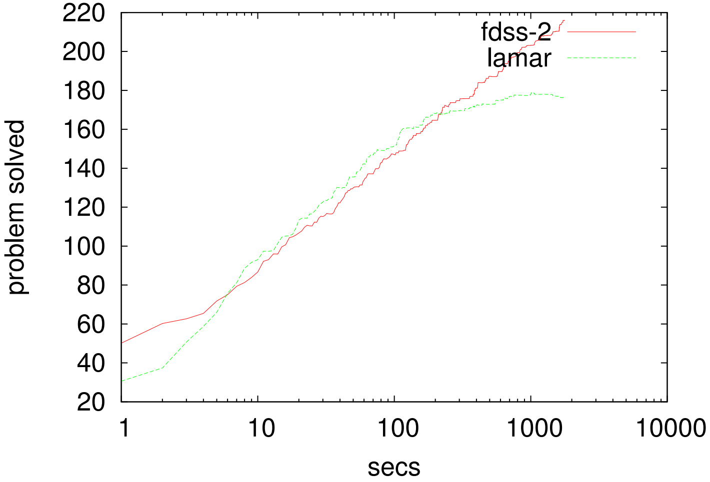


Figure: Evolution of coverage over time, sequential satisficing track IPC-2011.

Evaluation setup

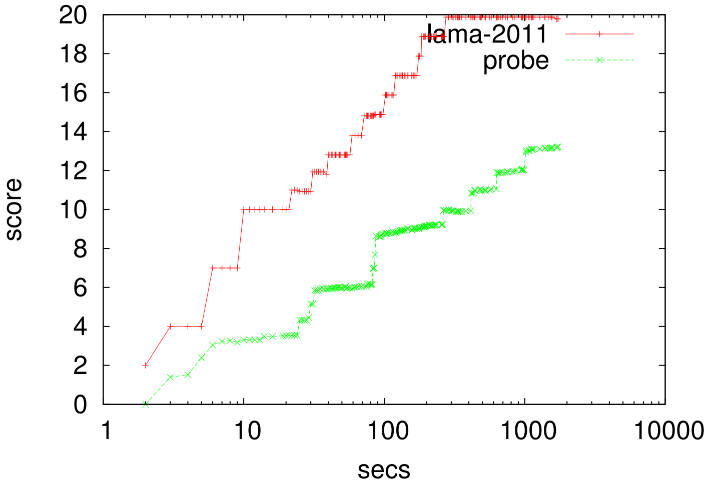


Figure: Evolution of the IPC score over time at the *openstacks* domain sequential satisficing track IPC-2011.

Evaluation setup

planner	Time failures	Memory failures	Unexpected failures
CPT4	176	0	56
GAMER	65	39	26
LMFORK	123	8	1
FD-AUTOTUNE	111	3	–
LMCUT	110	3	–
FORKINIT	87	33	2
SELMAX	100	9	2
FDSS-1	95	0	–
FDSS-2	3	78	17
IFORKINIT	58	71	7
BJOLP	29	81	19
MERGE-AND-SHRINK	4	76	31

Table: Number of *time*, *memory* and *unexpected* failures at the sequential optimal track of the IPC-2011.

Evaluation setup

domain	solved
visitall	20.00
transport	20.00
woodworking	19.00
scanalyzer	17.00
pegsol	15.00
parcprinter	13.00
barman	12.00
nomystery	10.00
floortile	8.00
parking	3.00
tidybot	0.00
elevators	0.00
openstacks	0.00
sokoban	0.00
total	**137.00**

domain	solved
visitall	20.00
transport	20.00
woodworking	19.00
scanalyzer	19.00
parking	18.00
barman	15.00
parcprinter	13.00
pegsol	12.00
nomystery	12.00
floortile	7.00
tidybot	0.00
elevators	0.00
openstacks	0.00
sokoban	0.00
total	**155.00**

Table: Problems solved by *yahsp2-mt* at the sequential satisficing track and at the sequential multicore track (**4 cores**) IPC-2011.

Evaluation setup

- Score function
- Computational resources
- Domains/problems

Evaluation setup

planner	nomystery	elevators	floortile	total
fd-autotune-2	18.36	16.17	8.87	**43.40**
arvand	18.97	11.22	3.00	**33.19**
forkuniform	10.45	18.01	4.02	**32.48**
fdss-2	11.21	14.50	6.60	**32.31**
fdss-1	11.26	12.52	5.30	**29.08**
fd-autotune-1	9.50	11.04	5.46	**26.00**
lama-2011	9.92	10.28	5.49	**25.69**
roamer	9.67	13.61	2.38	**25.65**
brt	5.75	13.84	2.82	**22.41**
lamar	11.46	7.34	2.36	**21.15**
lama-2008	11.44	4.94	2.07	**18.45**
probe	5.90	8.24	2.83	**16.98**
cpt4	15.00	0.00	0.00	**15.00**
randward	8.55	4.29	2.00	**14.84**
daeyahsp	9.67	0.00	4.39	**14.06**
madagascar-p	13.93	0.00	0.00	**13.93**
yahsp2-mt	9.61	0.00	4.08	**13.69**
popf2	8.22	4.73	0.67	**13.61**
madagascar	12.98	0.00	0.00	**12.98**
lprpgp	7.26	4.56	1.09	**12.90**
cbp2	4.00	7.34	0.00	**11.34**
yahsp2	6.70	0.00	3.29	**9.99**
cbp	4.00	4.86	0.00	**8.86**
satplanlm-c	3.00	0.00	0.00	**3.00**
sharaabi	0.00	0.56	0.00	**0.56**
acoplan	0.00	0.00	0.00	**0.00**
total	**236.79**	**168.04**	**66.73**	

Table: Score in a biased selection of domains from the seq-sat track IPC-2011.

Planning task

Beyond syntax, structural information affects planning performance

- Classical planning [Hoffmann, 2005]
  - goals dependencies, dead-ends,. . .

Evaluating planners

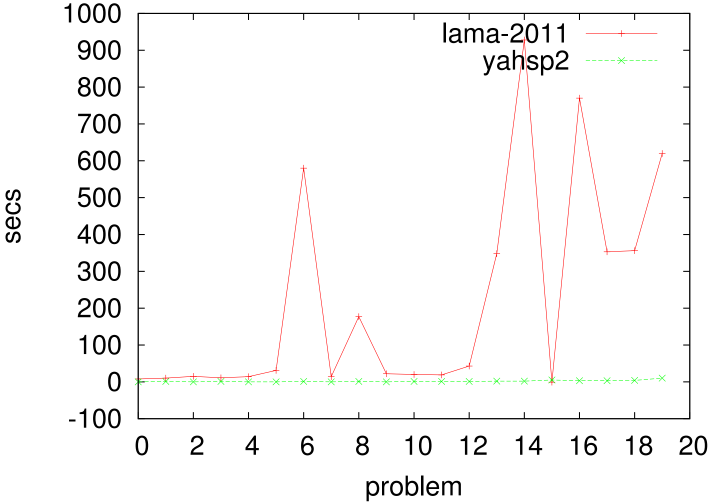


Figure: Time first solution *Transport* domain sequential satisficing track IPC-2011.

Planning task

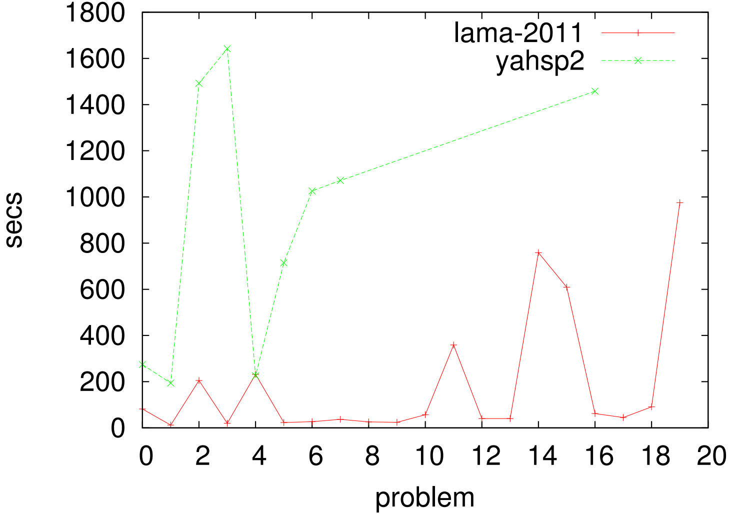


Figure: Time first solution *Parking* domain sequential satisficing track IPC-2011.

Evaluation setup

	lama-2011	yahsp2
000	1.00	0.56
001	1.00	0.65
002	1.00	0.54
003	1.00	0.56
004	1.00	0.86
005	1.00	0.58
006	1.00	0.72
007	1.00	0.60
008	1.00	$\varnothing$
009	1.00	$\varnothing$
010	1.00	$\varnothing$
011	1.00	$\varnothing$
012	1.00	$\varnothing$
013	1.00	$\varnothing$
014	1.00	$\varnothing$
015	1.00	$\varnothing$
016	1.00	0.60
017	1.00	$\varnothing$
018	1.00	$\varnothing$
019	1.00	$\varnothing$
total	**20.00**	**5.66**

Table: Score in the problems from the *parking* domain of the seq-sat track IPC-2011.

Planning task

Beyond syntax, structural information affects planning performance

- Classical planning [Hoffmann, 2005]
  - goals dependencies, dead-ends
- Temporal planning [Cushing et al., 2007]
  - Required concurrency

Planning task

planner	pegsol	crewp	parking	ostacks	elevat.	ftile	mcellar	sokoban	storage	pprinter	t&o	tms	total
dae-yahsp	19.67	19.95	18.92	20.00	14.46	7.96	0.00	4.55	17.06	3.58	0.00	0.00	**126.16**
yahsp2-mt	17.77	15.93	15.44	12.18	11.73	9.54	0.00	11.83	8.86	7.85	0.00	0.00	**111.14**
popf2	18.61	20.00	17.98	15.19	2.20	0.00	19.99	2.63	0.00	0.00	9.00	5.00	**110.60**
yahsp2	16.96	15.97	13.44	12.74	11.35	7.78	0.00	11.14	2.74	6.85	0.00	0.00	**98.97**
lmtd	19.95	0.00	0.00	0.00	7.73	5.00	15.00	0.00	0.00	0.00	10.07	0.00	**57.75**
cpt4	18.67	7.00	0.00	0.00	0.00	13.74	0.00	0.00	0.00	5.00	0.00	0.00	**44.41**
sharaabi	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**
tlp-gp	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.00**
total	**111.63**	**78.85**	**65.79**	**60.11**	**47.48**	**44.03**	**34.99**	**30.15**	**28.66**	**23.29**	**19.07**	**5.00**	

Table: Final scores temporal satisficing track IPC-2011.

Summary

Which planner should I buy?

- Planning task
  - Planning model
  - Performance metric
  - Benchmark
- Evaluation setup
  - Score function
  - Computational resources
  - Domains/problems

Evaluating planners

IPC-style experiments is a tradition [Hoffmann, 2011]

1. Run IPC benchmarks (unless you run all, run the most recent ones)
2. Time-out is 30 minutes
3. VALidate solutions [Howey et al., 2004]
4. Compare to the most recent IPC winner (using IPC score)

Outline

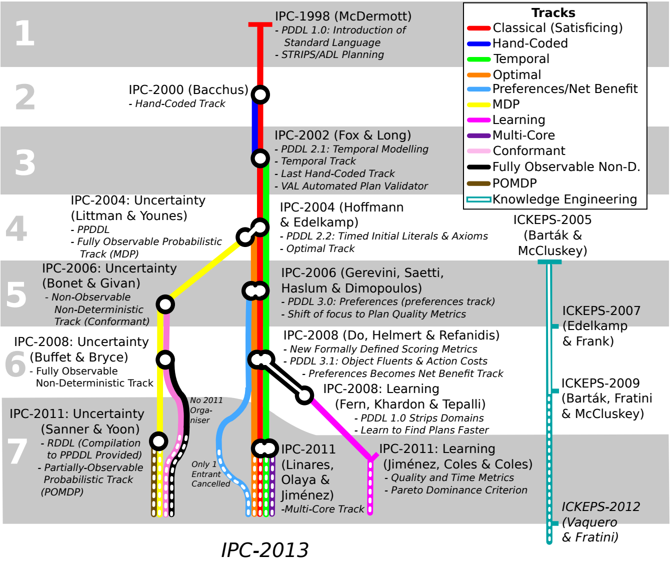
1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. Statistical Tests
5. Evaluation reports
6. Homework

IPC Evaluation

- Well-defined planning tasks
- Well-defined evaluation setup
- Available open-source tools

IPC Evaluation

History of the International Planning Competition



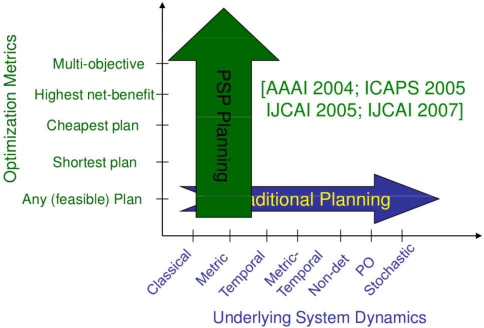
IPC Evaluation

Well-defined planning tasks,

- separation of
  - domain-**dependent** and domain-**independent** planning [Bacchus, 2001, Long and Fox, 2003a]
  - **optimal** and **satisfycing** planning [Hoffmann and Edelkamp, 2005]
  - **plan-cost** and **makespan** optimization `http://raos-ruminations.blogspot.com`
- 4 separated tracks at the IPC-2011
  - *seq-sat*, *seq-opt*, *seq-mco*, *tempo-sat* `http://www.plg.inf.uc3m.es/ipc2011-deterministic/`

IPC Evaluation

. . . but there is a multitude of different planning tasks not addressed at IPC [Kambhampati, 2011]



IPC Evaluation

. . . evenmore, planners at IPC are not implementing the full PDDL

PDDL	Requirements	Satisf.	Optimal	Multi-core	Temporal
1.2	typed representations	27	12	8	8
1.2	untyped representations	21	12	7	7
1.2	schematic representations	27	12	7	8
1.2	grounded representations	23	1	7	6
1.2	negative conditions	16	1	6	0
1.2	ADL conditions	15	1	6	1
1.2	conditional effects	15	0	5	1
1.2	universal effects	18	1	5	2
2.2	derived predicates	11	0	3	0
2.2	time-initial literals	–	–	–	3
3.1	numeric state variables	–	–	–	3
3.1	object fluent representations	0	0	0	0
Total		27	12	8	8

Table: PDDL coverage of the competing planners at the different tracks of IPC-2011.

IPC Evaluation

. . . skipping interesting domains

Domain	Authors	Remarks
Crisp	Ron Petrick and Alexander Koller	Required conditional-effects and quantified-preconditions which were not supported by most of participant planners
Market	Amanda Coles and Andrew Coles	Required numeric preconditions which were not supported by most of participant planners
Contingent Domains	Guy Shani	A collection of contingent planning domains compiled into classical planning. Required conditional-effects and quantified-preconditions which were not supported by most of participant planners

Table: Interesting domains out of IPC-2011, more info can be found at `http://www.plg.inf.uc3m.es/ipc2011-deterministic/NonUsedDomains`.

IPC Evaluation

From virtue to vice

- The IPC is a standard evaluation for a set of planning tasks but not for anything else
    - time-line based planning
    - model-lite planning
    - continuous planning
    - ...

IPC Evaluation

The *expressiveness* vs *performance* tension

- There is a lack of expressive planners at the IPC
- Classical planners can be used for further planning tasks [Nebel, 2000, Keyder and Geffner, 2009, Palacios and Geffner, 2009, Nguyen et al., 2012a]

IPC Evaluation

- Well-defined planning tasks
- Well-defined evaluation setup
- Available open-source tools

IPC Evaluation

Well-defined evaluation setup

- Score
- Computational resources
- Domains/Problems

IPC Evaluation

Once again there are interesting challenges out of the IPC evaluation setup,

- planning with small time bounds (videogames, robotics)
- efficient preprocessing (large logistics problems)
- using the Graphics Processing Unit (GPU) [Sulewski et al., 2011]
- using external memory [Edelkamp et al., 2007]
- ...

IPC Evaluation

From virtue to vice,

- the IPC is not an analysis of the current state-of-the-art but influences the shape of state-of-the-art planners
  - planners perform well on past IPC benchmarks
  - proliferation of portfolios and auto-tuned planners
  - planners tuned for the IPC evaluation setup

IPC Evaluation

- Well-defined planning tasks
- Well-defined evaluation setup
- Available open-source tools

IPC Evaluation

Available open-source tools to

- VALidate plans and reported metrics [Howey et al., 2004]
- share domains/problems/results
- run IPC-style experiments
- inspect results
- rank planners according to different metrics
- perform statistical tests

Summary

- Well defined evaluations
- Useful open-source software
- IPC evaluates a few interesting challenges not all of them

Outline

1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. **Statistical Tests**
5. Evaluation reports
6. Homework

The need of statistical tests (I)

- Overall, we have to assess on the performance of incomplete algorithms!
- where there are a number of different metrics
- First-order statistical measures such as the mean, median are not sufficient (see Jöerg Hoffman, *Evaluating planning algorithms*)
- Even if you accompany of second-order statistical measures such as the variance, they are still incomplete —but admittedly better informed

Statistical Tests

The need of statistical tests (II)

- Example: toss a coin ten times, observe eight heads. Is the coin fair (i.e., what is its long run behavior?) and what is your residual uncertainty?
- You say, "If the coin were fair, then eight or more heads is pretty unlikely, so I think the coin isnt fair"
- Like proof by contradiction: Assert the opposite (the coin is fair) show that the sample result ($\leq$ 8 heads) has low probability p, reject the assertion, with residual uncertainty related to p
- Estimate p with a sampling distribution

Statistical Tests

Statistical Tests in planning [Linares López et al., 2013]

- Parametric vs non-parametric
- Data: nominal/categorical, (discrete/continuous) dichotomous, ordinal, interval or ratio
- Purposes:
  - *Coverage*: **Binomial Test**
  - *Time, memory and cost*
    - Paired or related samples: **Wilcoxon signed-rank test**
    - Unrelated or non-paired samples: **Mann-Whitney U Test**
  - *Ranking*: **Spearman rank-order correlation coefficient** r_s
- Available under many languages including Python and R

Statistical Tests

General procedure [Corder and Foreman, 2009]

1. State the Null (H_0) and Research Hypothesis
2. Set the level of risk α
3. Choose the appropriate test
4. Compute the test statistic
5. Determine the value needed for rejection of the Null Hypothesis
6. Compare the obtained value to the critical value
7. Interpret the results
8. Report the results

Statistical Tests

Binomial Test:

- It is an exact two-tailed sign test used with dichotomous data
- It provides statistical significance of the Null Hypothesis that both categories are equally likely to occur
- This test was selected by Hoffmann and Nebel [Hoffmann and Nebel, 2001] to provide statistical evidence that their planner, FF, performed significantly better with some collections of enhancements than with others
- Use it in ablation studies or to analyze coverage

Wilcoxon signed rank test (I)

- It is a two-tailed nonparametric statistical procedure for comparing two samples that are paired, or related
- It tests the Null Hypothesis that both samples come from the same distribution
- It uses the signed ranks as the positive and negative differences

$$\sum R_+ \quad \sum R_-$$

Wilcoxon signed rank test (II)

- It has been already used to compare the performance of planners with respect to speed and quality in the analysis of results of the third and fifth International Planning Competitions [Long and Fox, 2003b, Gerevini et al., 2009]
- Use it to compare the performance of two different planners with regard to the same set of planning instances

Mann-Whitney U tests (I)

- It compares two samples that are independent, or not related
- It assesses the Alternate Hypothesis that one of two samples of independent observations tends to have larger values than the other
- It combines and ranks both samples and assesses the probability that there is a random walk in the resulting rank

Mann-Whitney U tests (II)

- Use it to compare performance of a planner with regard to problems in different domains

Spearman rank-order correlation coefficient (I)

- It measures the relationship between two variables on an ordinal scale of measurement
- It tests the Null Hypothesis that the samples are not correlated
- It uses the following formula in the absence of ties

$$r_s = 1 - \frac{6 \sum D_i^2}{n(n^2 - 1)}$$

Spearman rank-order correlation coefficient (II)

- or use the following formula in the presence of ties

$$r_s = \frac{(n^3 - n) - 6 \sum D_i^2 - (T_x + T_y)/2}{\sqrt{(n^3 - n)^2 - (T_x + T_y)(n^3 - n) + T_x T_y}}$$

- Use it to compare different rankings (e. g., according to different metrics)

Outline

1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. Statistical Tests
5. Evaluation reports
6. Homework

Evaluating planners

It's about understanding the world
Not about "my apple flies faster than yours"

Jörg Hoffmann
(ICAPS 2011 Summer School)

We fail more often because we solve the wrong problem than because we get the wrong solution to the right problem

Russell Ackoff

Evaluation reports

Controlling complexity is the essence of computer programming

Brian Kernigan

Create simple (hopefully beautiful) and easy to understood views of your data . . .

Simplicity does not precede complexity, but follows it

Alan Perlis

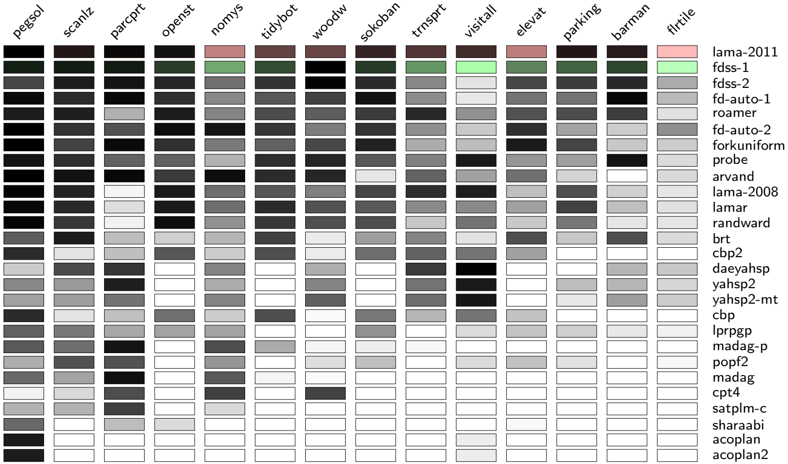
. . . it will help you understand the complex

Beauty is the ultimate defense against complexity

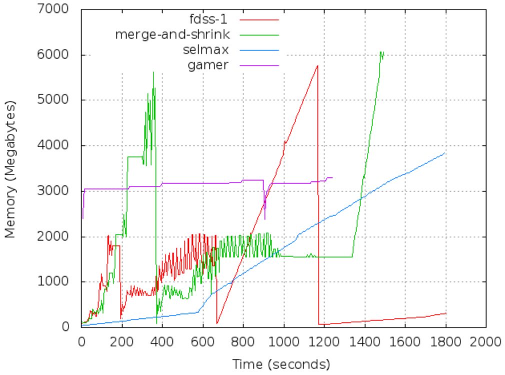
David Galernter

Evaluation reports

Sequential Satisficing track: Results

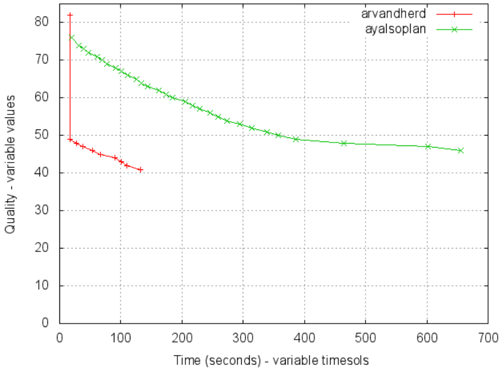


Evaluation reports



Memory profile of FDSS-1, MERGE-AND-SHRINK, SELMAX and GAMER for solving problem 018 of the domain WOODWORKING

Evaluation reports



Time (in seconds) when each solution file was generated and the value of the metric of the plans found by ARVANDHERD and AYALSOPLAN in problem 010 of the domain OPENSTACKS of the sequential multi-core track

Evaluation reports

- Failures: time, memory and *unexplained*
- Performance comparison
  - Fixed-time comparisons: total order (*ranking*) vs. partial order
  - Comparisons over time: probability distributions and score landscapes
- Comparing performance improvement

Evaluation reports

- Failures: time, memory and *unexplained*
- Performance comparison
  - Fixed-time comparisons: total order (*ranking*) vs. partial order
  - Comparisons over time: probability distributions and score landscapes
- Comparing performance improvement

Evaluation reports

. . . but the devil is in the details!

Jörg Hoffmann
(ICAPS 2011 Summer School)

Beware of the man who won't be bothered with details

William Feather, Sr.

Evaluation reports

	LAMA-2011	FDSS-1	FDSS-2	FD-AUTOTUNE-1	ROAMER
Score	216.33	202.08	196.00	185.09	181.47
Solved	250	232	233	223	213
Success ratio	89.28%	82.85%	83.21%	79.64%	76.07%

	FD-AUTOTUNE-2	FORKUNIFORM	PROBE	ARVAND	LAMA-2008
Score	178.15	177.91	177.14	165.07	163.33
Solved	193	207	233	190	188
Success ratio	68.92%	73.92%	83.21%	67.85%	67.14%

	LAMAR	RANDWARD	BRT	CBP2	DAE_YAHSP
Score	159.20	141.43	116.01	98.34	95.23
Solved	195	184	157	135	110
Success ratio	69.64%	65.71%	56.07%	42.85%	39.28%

	YAHSP2	YAHSP2-MT	CBP	LPRPGP	MADAGASCAR-P
Score	94.97	90.95	85.43	67.07	65.93
Solved	138	132	123	118	88
Success ratio	49.28%	47.14%	43.92%	42.14%	31.42%

	POPF2	MADAGASCAR	CPT4	SATPLANLM-C	SHARAABI
Score	59.88	51.98	47.85	29.96	20.52
Solved	81	67	52	32	33
Success ratio	28.92%	23.92%	18.57%	11.42%	11.78%

...

Evaluation reports

What happened to Mp? (I)

domain	oknumsolved	numtimefails	nummemfails	numunexfails
barman	0	20	0	0
elevators	0	0	0	2
floortile	0	0	0	20
nomystery	15	1	0	4
openstacks	0	10	0	10
parcprinter	20	0	0	0
parking	0	20	0	0
pegsol	20	0	0	0
scanalyzer	18	0	0	2
sokoban	2	18	0	0
tidybot	10	2	0	8
transport	2	11	2	5
visitall	0	20	0	0
woodworking	1	0	0	0

Evaluation reports

What happened to Mp? (& II)

domain	numsolved	oknumsolved	numtimefails	nummemfails	numunexfails
barman	0	0	20	0	0
elevators	18	0	0	0	2
floortile	0	0	0	0	20
nomystery	15	15	1	0	4
openstacks	0	0	10	0	10
parcprinter	20	20	0	0	0
parking	0	0	20	0	0
pegsol	20	20	0	0	0
scanalyzer	18	18	0	0	2
sokoban	2	2	18	0	0
tidybot	10	10	2	0	8
transport	2	2	11	2	5
visitall	0	0	20	0	0
woodworking	20	1	0	0	0

Evaluation reports

There was a bug!

Domain	M	Mp
barman	0 / 0	0 / 0
elevators	1 / 0	19 / 0
floortile	20 / 0	20 / 0
nomystery	15 / 17	15 / 15
openstacks	0 / 0	0 / 0
parcprinter	20 / 20	20 / 20
parking	0 / 0	0 / 0
pegsol	17 / 17	20 / 20
scanalyzer	12 / 11	18 / 18
sokoban	0 / 0	2 / 2
tidybot	0 / 1	12 / 10
transport	0 / 0	2 / 2
visitall	0 / 0	0 / 0
woodworking	20 / 1	20 / 1
Total	105 / 67	148 / 88

Evaluation reports

- Failures: time, memory and *unexplained*
- Performance comparison
  - Fixed-time comparisons: total order (*ranking*) vs. partial order
  - Comparisons over time: probability distributions and score landscapes
- Comparing performance improvement

Evaluation reports

	LAMA-2011	FDSS-1	FDSS-2	FD-AUTOTUNE-1	ROAMER
Score	216.33	202.08	196.00	185.09	181.47
Solved	250	232	233	223	213
Success ratio	89.28%	82.85%	83.21%	79.64%	76.07%
	FD-AUTOTUNE-2	FORKUNIFORM	PROBE	ARVAND	LAMA-2008
Score	178.15	177.91	177.14	165.07	163.33
Solved	193	207	233	190	188
Success ratio	68.92%	73.92%	83.21%	67.85%	67.14%

...

Evaluation reports

	LAMA-2011	FDSS-1	FDSS-2	FD-AUTOTUNE-1	ROAMER
Score	216.33	202.08	196.00	185.09	181.47
Coverage	250	232	233	223	213
Time	155.27	99.63	137.26	129.59	118.81
QT	207.98	163.73	180.79	172.65	170.38

	FD-AUTOTUNE-2	FORKUNIFORM	PROBE	ARVAND	LAMA-2008
Score	178.15	177.91	177.14	165.07	163.33
Coverage	193	207	233	190	188
Time	103.84	113.67	154.74	77.46	101.76
QT	151.96	158.11	185.35	137.74	151.98

...

Evaluation reports

- Be concise!
- Formulate a hypothesis:

  $\mathbf{H_0}$: *Score is correlated with the other metrics*

- and choose a confidence level:

  $$\alpha = 0.999$$

- In this case, the Spearman rank-order correlation coefficient r_s will test this hypothesis —without assuming any underlying distribution

Evaluation reports

	Coverage	Time	QT
Score	0.974 0.000	0.893 0.000	0.969 0.000
Coverage		0.945 0.000	0.992 0.000
Time			0.956 0.000

The Spearman rank-order correlation coefficient r_s is shown above and the two-tailed significance p is shown below.

The hypothesis is accepted!

Evaluation reports

	ARVANDHERD	AYALSOPLAN	PHSFF	ROAMER-P	YAHSP2-MT
Score	227.07	159.95	130.59	129.06	118.58
Solved	236	184	163	140	153
Success ratio	84.28%	65.71%	58.21%	50.0%	54.64%

	MADAGASCAR-P	MADAGASCAR	ACOPLAN
Score	66.44	52.00	17.62
Solved	88	67	18
Success ratio	31.42%	23.92%	6.42%

Official results of the IPC 2011 sequential multi-core track

Evaluation reports

Loot at raw data! Not only at summaries!

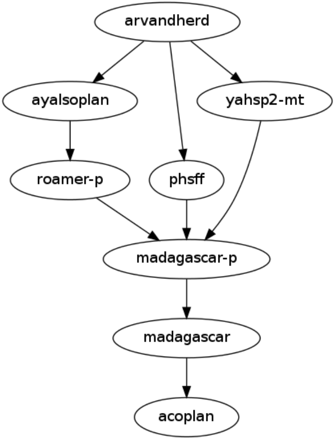
Jörg Hoffmann
(ICAPS 2011 Summer School)

Be aware there might be automated means to do it!

The enjoyment of one's tools is an essential ingredient of successful work

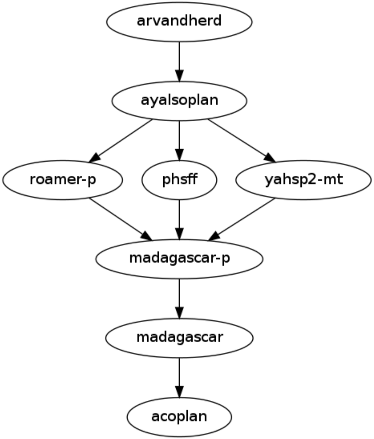
Donald E. Knuth

Evaluation reports



Partial order of the performance of planners in the sequential multi-core track in terms of *successfully solved problems* according to the Binomial test with $p = 0.5$. The statistical significance is 99.9%

Evaluation reports



Partial order of the performance of planners in the sequential multi-core track in terms of *quality* according to the Wilcoxon signed-rank test.
The statistical significance is 99.9%

Evaluation reports

- Failures: time, memory and *unexplained*
- Performance comparison
  - Fixed-time comparisons: total order (*ranking*) vs. partial order
  - Comparisons over time: probability distributions and score landscapes
- Comparing performance improvement

Evaluation reports

A PhD student (you!) and your PhD advisor are having a discussion about two algorithms:

- I used 61 problems from the *Blocksworld* domain. The first algorithm solves 51 problems and the second one solves 58. So it seems that the second algorithm is better
- Better for what?
- Well, I was assuming coverage
- Hmmm, . . . , that's unclear but what about time?
- Oh, no prob, I also realized that the second algorithm is faster
- Really?
- Well ...

Evaluation reports

Cut-offs (*such as time*) may bias the sample!

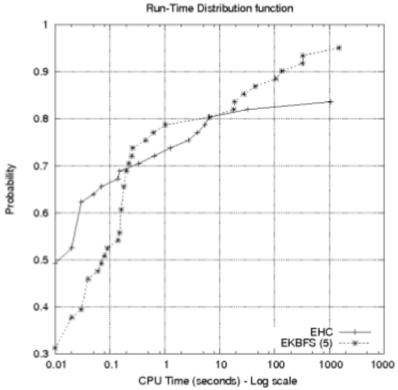
Jörg Hoffmann
(ICAPS 2011 Summer School)

Definition

Consider a heuristic algorithm A for solving a finite and known set of problems in the planning domain $\mathcal{D}$, and let $P(RT_{A,\mathcal{D}} \leq t)$ denote the probability that A finds a solution for one of these instances in time less or equal than t. The Run-Time Distribution (or RTD, for short) of A on $\mathcal{D}$ is the probability distribution of the random variable $RT_{A,\mathcal{D}}$, which is characterized by the Run-Time Distribution function $rtd : \mathbb{R}^+ \mapsto [0,1]$ defined as
$rtd(t) = P(RT_{A,\mathcal{D}} \leq t)$

Used since 2005 but very scarcely [Haslum et al., 2005]

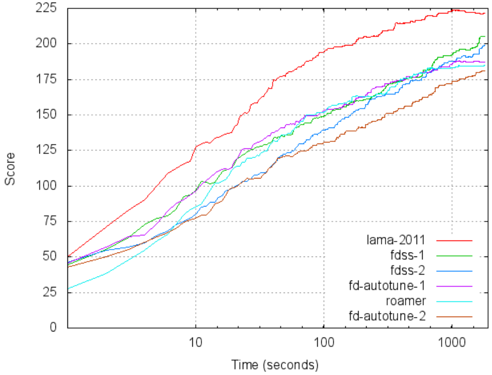
Evaluation reports



So it seems that below $t = 0.2$ seconds, EHC is significantly more effective and though EKBFS (5) is better in the *long-term*, they are more or less equivalent again around $t = 10$

Clearly, EKBFS (5) has better overall coverage than EHC

Evaluation reports



Evolution of the metric *quality* over time for the first six planners of the sequential satisficing track of the seventh International Planning Competition

Evaluation reports

- Failures: time, memory and *unexplained*
- Performance comparison
  - Fixed-time comparisons: total order (*ranking*) vs. partial order
  - Comparisons over time: probability distributions and score landscapes
- Comparing performance improvement

Evaluation reports

Planner	**Base performance**	**DSK performance**
A	10	15
B	5	10

$Delta\ performance = (DSK\ performance - Base\ performance)$ is clearly insufficient

> *Branching is easy. Merging is hard*
>
> *Eric Sink*

Evaluation reports

Comparing performance improvement

- There are cases where one is interested in two- (or multi-) variate analysis
- This need arises often in the learning and multi-core tracks, but also in others
- It is relevant, for example, to consider representational issues such as the impact of macro-actions and entanglements

One alternative are ablation studies (see Jöerg Hoffman, *Evaluating planning algorithms*)

Evaluation reports

Definition

qt computes for each planner and task a tuple (Q, T) where Q stands for the quality of the best solution found by the same planner and T is the time it took for the planner to find it. Next, it awards each planner with a score that equals the number of tuples it pareto-dominates

Definition

(Q, T) is said to pareto-dominate (Q', T') if and only if $Q \leq Q'$ and $T \leq T'$

Evaluation reports

planner	pegsol	crwpln	prking	opstcks	elvtrs	flrtle	matchc	skban	storage	prcprnt	t&o	tms	total
dae_yahsp	19.67	19.95	18.92	20.00	14.46	7.96	0.00	4.55	17.06	3.58	0.00	0.00	**126.16**
yahsp2-mt	17.77	15.93	15.44	12.18	11.73	9.54	0.00	11.83	8.86	7.85	0.00	0.00	**111.14**
popf2	18.61	20.00	17.98	15.19	2.20	0.00	19.99	2.63	0.00	0.00	9.00	5.00	**110.60**
yahsp2	16.96	15.97	13.44	12.74	11.35	7.78	0.00	11.14	2.74	6.85	0.00	0.00	**98.97**
total	**111.63**	**78.85**	**65.79**	**60.11**	**47.48**	**44.03**	**34.99**	**30.15**	**28.66**	**23.29**	**19.07**	**5.00**	Quality

planner	pegsol	crwpln	prking	opstcks	elvtrs	flrtle	matchc	skban	storage	prcprnt	t&o	tms	total
yahsp2-mt	16.63	18.43	19.00	19.00	18.61	11.95	0.00	11.00	12.00	8.00	0.00	0.00	**134.62**
yahsp2	15.97	18.27	20.00	17.44	18.42	9.78	0.00	4.50	10.43	7.00	0.00	0.00	**121.80**
dae_yahsp	17.17	19.03	19.80	16.84	13.87	8.33	0.00	17.93	4.07	3.86	0.00	0.00	**120.89**
popf2	16.42	19.60	19.80	17.78	2.60	0.00	20.00	0.00	1.98	0.00	9.00	5.00	**112.17**
total	**100.48**	**82.13**	**78.60**	**71.06**	**61.64**	**49.15**	**35.00**	**33.43**	**28.48**	**23.86**	**20.71**	**5.00**	QT

planner	pegsol	crwpln	prking	opstcks	elvtrs	flrtle	matchc	skban	storage	prcprnt	t&o	tms	total
yahsp2-mt	19.68	20.00	19.00	18.85	15.69	12.17	10.83	0.00	12.00	8.00	0.00	0.00	**136.23**
yahsp2	20.00	18.35	20.00	18.99	18.70	9.62	4.32	0.00	7.72	7.00	0.00	0.00	**124.69**
popf2	16.35	15.43	11.36	9.24	1.26	0.00	0.00	20.00	1.67	0.00	9.00	5.00	**89.31**
dae_yahsp	17.80	15.49	5.19	5.90	4.66	4.07	13.64	0.00	2.21	2.70	0.00	0.00	**71.65**
total	**107.19**	**75.76**	**55.56**	**52.98**	**43.56**	**43.11**	**28.79**	**27.16**	**23.60**	**21.95**	**16.85**	**5.00**	Time

- Analyze source of failures

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)
- Use one in your toolbox: Identify a suitable report

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)
- Use one in your toolbox: Identify a suitable report
- Use summaries, but dig also into raw data

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)
- Use one in your toolbox: Identify a suitable report
- Use summaries, but dig also into raw data
- Does it solve your question? If not, start again

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)
- Use one in your toolbox: Identify a suitable report
- Use summaries, but dig also into raw data
- Does it solve your question? If not, start again
- If yes, do the answer post additional questions? If yes, start again

Evaluation reports · Summary

- Analyze source of failures
- Look at your data, identify the right problem
- Make a hypothesis (as a positive statement)
- Use one in your toolbox: Identify a suitable report
- Use summaries, but dig also into raw data
- Does it solve your question? If not, start again
- If yes, do the answer post additional questions? If yes, start again
- If not, start again anyway!

Evaluation reports · Summary

To err is human, but to really foul things up you need a computer

Paul Ehrlich

...and also the other way round!

Computer science is no more about computers than astronomy is about telescopes

Edsger W. Dijkstra

Do good implementations, but get rid of improving your results with technical tricks

Evaluation reports · Summary

In general you [become successful] not by knowing what the experts know but by learning what they think is beneath them

George Gilder

Imitate others but do not do the same thing!
Remember, it is about understanding the world!

Evaluation reports · Summary

Somewhere, something incredible is waiting to be known

Carl Sagan

Overall, be curious!

Outline

1. Planning task
2. Evaluation setup
3. IPC Evaluation
4. Statistical Tests
5. Evaluation reports
6. Homework

Homework

Questions

1. (seq-opt) Report the number of memory, time and unexplained failures of every entrant
2. (tempo-sat) How many problems were solved by YAHSP2 and how many were valid? Show the results per domain
3. (seq-sat) How long did it take FDSS-1 to find the first and last solution in each problem of the domain TRANSPORT?
4. (seq-opt) Show the final score of every entrant according to the official metric of the IPC 2011
5. (seq-mco) Show the progress of coverage for the planners ARVANDHERD and AYALSOPLAN

Homework

Challenges

1. (tempo-sat) Which planner (among those solving at least 1 problem) show the highest ratio of invalid plan solution files? What domains were harder for that planner?
2. (seq-mco) In what domain do ARVANDHERD achieves full coverage faster?
3. (seq-sat) Create a figure that shows the difference between the best and worst plan quality found by FDSS-2 as a function of the time to find them in domain OPENSTACKS
4. (tempo-sat) Show the progress of plan cost and plan length of all the solutions found by YAHSP2-MT in problem 003 of domain CREWPLANNING
5. (seq-opt) Compare the results of a statistical test on plan quality with $\alpha = 0.005$ and $\alpha = 0.001$

Bibliography I

Bacchus, F. (2001).
AIPS 2000 planning competition: The fifth international conference on artificial intelligence planning and scheduling systems.
AI Magazine, 22(3):47–56.

Brafman, R. I. and Domshlak, C. (2013).
On the complexity of planning for agent teams and its implications for single agent planning.
Artif. Intell., 198:52–71.

Corder, G. W. and Foreman, D. I. (2009).
Nonparametric Statistics for Non-Statisticians.
John Wiley & Sons, New Jersey, United States.

Cushing, W., Kambhampati, S., Mausam, and Weld, D. S. (2007).
When is temporal planning really temporal?
In *Proceedings of the 20th international joint conference on Artifical intelligence*, IJCAI'07, pages 1852–1859.

Edelkamp, S., Jabbar, S., and Bonet, B. (2007).
External memory value iteration.
In *Proceedings of the Seventeenth International Conference on Automated Planning and Scheduling, ICAPS 2007*, pages 128–135.

Bibliography II

Fox, M., Gerevini, A., Long, D., and Serina, I. (2006).
Plan stability: Replanning versus plan repair.
In *In Proc. ICAPS*, pages 212–221. AAAI Press.

Gerevini, A. E., Haslum, P., Long, D., Saetti, A., and Dimopoulos, Y. (2009).
Deterministic planning in the fifth international planning competition: PDDL3 and experimental evaluation of the planners.
Artificial Intelligence, 173(5-6):619–668.

Haslum, P. (2012).
Narrative planning: Compilations to classical planning.
J. Artif. Intell. Res. (JAIR), 44:383–395.

Haslum, P., Bonet, B., and Geffner, H. (2005).
New admissible heuristics for domain-independent planning.
In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, pages 1163–1168, Pittsburgh, United States.

Hoffmann, J. (2005).
Where "ignoring delete lists" works: Local search topology in planning benchmarks.
Journal of Artificial Intelligence Research, 24:685–758.

Bibliography III

Hoffmann, J. (2011).
Evaluating planning algorithms, or: How to count sheep?
ACAI Summer School on Automated Planning and Scheduling, ICAPS 2011.

Hoffmann, J. and Edelkamp, S. (2005).
The deterministic part of IPC-4: An overview.
J. Artif. Intell. Res. (JAIR), 24:519–579.

Hoffmann, J. and Nebel, B. (2001).
The FF planning system: Fast plan generation through heuristic search.
Journal of Artificial Intelligence Research, 14:253–302.

Howe, A. E. and Dahlman, E. (2002).
A critical assessment of benchmark comparison in planning.
J. Artif. Intell. Res. (JAIR), 17:1–3.

Howey, R., Long, D., and Fox, M. (2004).
VAL: Automatic plan validation, continuous effects and mixed initiative planning using PDDL.
In *The Sixteenth IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2004)*, pages 294–301, Boca Raton, Florida, United States.

Bibliography IV

Kambhampati, S. (2011).
Back to the future of planning.
ACAI Summer School on Automated Planning and Scheduling, ICAPS 2011.

Keyder, E. and Geffner, H. (2009).
Soft goals can be compiled away.
Journal of Artificial Intelligence Research, 36:547–556.

Linares, C., Jimnez, S., and Helmert, M. (2013).
Automating the evaluation of planning systems.
AI Communications.

Linares López, C., Jiménez, S., and Helmert, M. (2013).
Automating the evaluation of planning systems.
AI Communications.

Long, D. and Fox, M. (2003a).
The 3rd international planning competition: Results and analysis.
J. Artif. Intell. Res. (JAIR), 20:1–59.

Long, D. and Fox, M. (2003b).
The 3rd international planning competition: Results and analysis.
Journal of Artificial Intelligence Research, 20:1–59.

Bibliography V

Nebel, B. (2000).
Logic-based artificial intelligence.
chapter On the expressive power of planning formalisms, pages 469–488.

Nguyen, H.-K., Tran, D.-V., Son, T. C., and Pontelli, E. (2012a).
On computing conformant plans using classical planners: A generate-and-complete approach.
pages 190–198, São Paulo, Brazil. AAAI.

Nguyen, T. A., Do, M. B., Gerevini, A., Serina, I., Srivastava, B., and Kambhampati, S. (2012b).
Generating diverse plans to handle unknown and partially known user preferences.
Artif. Intell., 190:1–31.

Nguyen, X. and Kambhampati, S. (2001).
Reviving partial order planning.
In *International Joint Conference on Artificial Intelligence*, pages 459–466.

Palacios, H. and Geffner, H. (2009).
Compiling uncertainty away in conformant planning problems with bounded width.
J. Artif. Intell. Res. (JAIR), 35:623–675.

Bibliography VI

Sulewski, D., Edelkamp, S., and Kissmann, P. (2011).
Exploiting the computational power of the graphics card: Optimal state space planning on the GPU.
pages 242–249, Freiburg, Germany.