

Purse-Based Scoring for Comparison of Exponential-Time Programs

Allen Van Gelder Daniel Le Berre
 University of California, Santa Cruz CRIL, Université d'Artois
 E-mail avg@cs.ucsc.edu. E-mail leberre@cril.univ-artois.fr

Armin Biere Oliver Kullmann Laurent Simon
 JKU, Linz, Austria University of Wales, Swansea LRI, Université Paris 11
 E-mail biere@jku.at E-mail O.Kullmann@swansea.ac.uk E-mail simon@lri.fr

June 17, 2005

Abstract

A purse-based method for scoring solving competitions is introduced. Its application is intended for benchmark suites in which it is expected that solvers will not be able to solve all instances. The main idea is that each benchmark problem has an associated purse (in the sense of prize) that is divided among those solvers that are able to solve it. There is no "penalty" for failing to solve an instance beyond not sharing in that purse. Properties of this scoring method are discussed. Preliminary experimental data is given, based on stage one of the satisfiability solver competition held in connection with SAT 2005, St. Andrews, Scotland, June 2005.

- It stabilizes the rankings of the solvers at the end of the competition.

While the scoring scheme was designed on a purely theoretical basis, the results of the SAT 2005 Competition indicate that the new scoring scheme meets its expectations in practice.

3 The Purse-Based Scoring System

The implemented scoring plan works as follows. A *run* is defined to be the execution of one *solver* on one benchmark instance, or *problem*. Each run is allocated a certain amount of CPU time. If the solver succeeds, *timeUsed* records the time.

For SAT 2005, there are three categories of benchmark, INDUSTRIAL, CRAFTED, and RANDOM. Within each category, there are several *specialties*, such as SAT, SAT+UNSAT, UNSAT, and CERTIFIED-UNSAT. The scoring system is applied separately within each combination of category and specialty.

Each problem has a *solution purse*, which is divided equally among all competition solvers that solve the problem. For SAT 2005, all problems have the standard solution purse (*stdP*).

Each problem has a *speed purse*, which is divided *unequally* among all competition solvers that solve the problem. The speed purse is a fixed multiple (*spdM*) of the solution purse for all problems in the entire competition; it gives a weighting between solving and speed.

The formula to divide the speed purse of a problem is the following, where *p* is problem-id and *s* and *i* are solvers, times are in seconds, and 10,000 is an arbitrary scale factor.

$$speedFactor(p, i) = \begin{cases} \frac{10000}{1 + timeUsed(p, i)} & \text{if } i \text{ solved } p; \\ 0 & \text{if } i \text{ did not solve } p. \end{cases} \quad (1)$$

$$speedAward(p, s) = \frac{speedPurse(p) * speedFactor(p, s)}{\sum_i speedFactor(p, i)} \quad (2)$$

Thus, the *speedAward* is pro rata by *speedFactor*.

The series purses reward breadth of application. Each series (within specialty within category) has a *series purse*, which is divided equally among all competition solvers that solve at least one problem in the series. If no solver solves any problem in a certain series, its series purse is not distributed.

For SAT 2005, all series containing 5 or more benchmark instances have the same series purse, which is a fixed multiple (*serM*) of the standard solution purse. (Recall that scoring is separately applied within each combination of category and specialty, e.g., SAT within RANDOM, or SAT+UNSAT within CRAFTED.) All series containing 4 or fewer benchmarks have the same series purse, which is a fixed multiple (*serM* / 3) of the standard solution purse.

The coefficients and multiples for SAT 2005 are:

$$stdP = 1000.0; \quad spdM = 1.0; \quad serM = 3.0.$$

4 Discussion

The new scoring scheme and particularly some of its parameters are a first shot. After the competition they most likely will need to be adjusted. The general goal should be to advance the state-of-the-art of SAT solvers. There are multiple contradictory interpretations what this means: speed on specific instances versus robustness versus breadth of application. We plan to investigate various intuitive parameter settings and compare the resulting ranking of the top solvers manually. It is hoped that only for extreme settings the ranking will change considerably. It is also important to verify that all scores have an influence on the final ranking. If a certain parameter is not important, its contribution is not needed and the scoring scheme can be simplified accordingly. In principle, it should be possible to adjust the parameters dynamically during at the next competition.

5 Preliminary Experimental Results

These tables present the results of stage one for the SAT 2005 Competition.

Table 1: INDUSTRIAL, best performers last.

| (A) SAT+UNSAT | | | | (B) SAT | | | | (C) UNSAT | | | |
|---------------|---------|------------|-------|----------|---------|------------|-------|-----------|---------|------------|-------|
| Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | |
| | | Sat | Unsat | | | Sat | Unsat | | | Sat | Unsat |
| solver36 | 1544.0 | 28 | 0 | solver36 | 1544.0 | 28 | 0 | solver1 | 0 | 0 | 0 |
| solver1 | 3154.0 | 50 | 0 | solver1 | 3154.0 | 50 | 0 | solver15 | 0 | 0 | 0 |
| solver43 | 3178.0 | 40 | 0 | solver43 | 3178.1 | 40 | 0 | solver24 | 0 | 0 | 0 |
| solver32 | 3442.2 | 56 | 0 | solver32 | 3442.2 | 56 | 0 | solver27 | 0 | 0 | 0 |
| solver24 | 4117.7 | 56 | 0 | solver24 | 4117.7 | 56 | 0 | solver28 | 0 | 0 | 0 |
| solver27 | 4563.3 | 59 | 0 | solver8 | 4422.4 | 61 | 0 | solver31 | 0 | 0 | 0 |
| solver28 | 5248.8 | 65 | 0 | solver7 | 4479.2 | 61 | 0 | solver32 | 0 | 0 | 0 |
| solver15 | 5291.4 | 65 | 0 | solver27 | 4563.3 | 59 | 0 | solver36 | 0 | 0 | 0 |
| solver42 | 5369.8 | 67 | 0 | solver9 | 4647.7 | 60 | 0 | solver42 | 0 | 0 | 0 |
| solver8 | 5530.1 | 60 | 5 | solver41 | 5175.8 | 67 | 0 | solver43 | 0 | 0 | 0 |
| solver9 | 5795.3 | 60 | 5 | solver28 | 5248.8 | 65 | 0 | solver8 | 907.7 | 0 | 5 |
| solver31 | 6028.8 | 71 | 0 | solver15 | 5291.4 | 65 | 0 | solver9 | 947.6 | 0 | 5 |
| solver7 | 7116.1 | 61 | 24 | solver42 | 5369.8 | 67 | 0 | solver38 | 994.4 | 0 | 5 |
| solver41 | 10051.7 | 67 | 39 | solver20 | 5960.7 | 66 | 0 | solver7 | 2436.8 | 0 | 24 |
| solver20 | 11820.4 | 66 | 29 | solver31 | 6028.8 | 71 | 0 | solver41 | 4576.1 | 0 | 39 |
| solver38 | 12623.2 | 82 | 5 | solver37 | 7080.9 | 80 | 0 | solver20 | 5836.1 | 0 | 29 |
| solver37 | 15271.8 | 80 | 63 | solver33 | 8445.5 | 88 | 0 | solver18 | 7393.9 | 0 | 45 |
| solver33 | 16793.1 | 88 | 61 | solver18 | 9809.3 | 91 | 0 | solver19 | 7411.3 | 0 | 45 |
| solver18 | 17645.9 | 91 | 45 | solver19 | 10739.3 | 91 | 0 | solver21 | 7480.8 | 0 | 65 |
| solver19 | 18593.2 | 91 | 45 | solver21 | 11262.1 | 94 | 0 | solver33 | 7919.3 | 0 | 61 |
| solver21 | 18885.5 | 94 | 65 | solver38 | 11605.2 | 82 | 0 | solver37 | 8266.2 | 0 | 63 |
| solver22 | 22358.2 | 96 | 65 | solver6 | 13416.4 | 94 | 0 | solver22 | 8420.5 | 0 | 65 |
| solver6 | 25106.7 | 94 | 78 | solver40 | 14290.2 | 91 | 0 | solver5 | 8470.4 | 0 | 67 |
| solver5 | 26955.2 | 88 | 67 | solver22 | 14598.7 | 96 | 0 | solver39 | 11631.3 | 0 | 60 |
| solver39 | 27201.3 | 90 | 60 | solver39 | 15350.1 | 90 | 0 | solver6 | 12684.7 | 0 | 78 |
| solver40 | 30400.5 | 91 | 78 | solver17 | 15821.4 | 94 | 0 | solver17 | 15324.4 | 0 | 77 |
| solver17 | 31312.2 | 94 | 77 | solver5 | 17756.5 | 88 | 0 | solver40 | 15985.5 | 0 | 78 |
| solver16 | 39359.3 | 99 | 79 | solver16 | 19051.5 | 99 | 0 | solver26 | 20029.2 | 0 | 74 |
| solver26 | 55638.8 | 114 | 74 | solver26 | 36651.5 | 114 | 0 | solver16 | 20802.1 | 0 | 79 |
| solver34 | 85602.9 | 117 | 78 | solver34 | 52497.0 | 117 | 0 | solver34 | 34481.2 | 0 | 78 |

Table 2: CRAFTED, best performers last.

| (A) SAT+UNSAT | | | | (B) SAT | | | | (C) UNSAT | | | |
|---------------|---------|------------|-------|----------|---------|------------|-------|-----------|---------|------------|-------|
| Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | |
| | | Sat | Unsat | | | Sat | Unsat | | | Sat | Unsat |
| solver1 | 5130.5 | 87 | 0 | solver36 | 1352.7 | 20 | 0 | solver1 | 0 | 0 | 0 |
| solver32 | 5221.3 | 97 | 0 | solver43 | 4205.0 | 49 | 0 | solver15 | 0 | 0 | 0 |
| solver28 | 7875.5 | 101 | 0 | solver1 | 5566.5 | 87 | 0 | solver27 | 0 | 0 | 0 |
| solver43 | 8269.1 | 49 | 34 | solver32 | 5735.2 | 96 | 0 | solver28 | 0 | 0 | 0 |
| solver27 | 9561.3 | 134 | 0 | solver38 | 7274.2 | 86 | 0 | solver31 | 0 | 0 | 0 |
| solver15 | 10393.5 | 128 | 0 | solver28 | 9420.5 | 99 | 0 | solver32 | 0 | 0 | 0 |
| solver31 | 11312.8 | 146 | 0 | solver27 | 10360.2 | 134 | 0 | solver42 | 0 | 0 | 0 |
| solver38 | 11912.3 | 86 | 50 | solver8 | 10543.0 | 125 | 0 | solver43 | 4249.4 | 0 | 35 |
| solver42 | 17047.5 | 132 | 0 | solver24 | 10837.4 | 61 | 0 | solver38 | 6000.2 | 0 | 50 |
| solver24 | 19055.6 | 61 | 37 | solver7 | 10848.5 | 127 | 0 | solver24 | 8551.5 | 0 | 36 |
| solver9 | 19868.0 | 124 | 63 | solver15 | 11060.1 | 127 | 0 | solver9 | 9902.5 | 0 | 63 |
| solver7 | 21063.6 | 127 | 66 | solver9 | 11392.7 | 124 | 0 | solver5 | 10613.6 | 0 | 83 |
| solver33 | 21179.4 | 136 | 73 | solver31 | 12184.3 | 146 | 0 | solver33 | 11204.4 | 0 | 73 |
| solver8 | 21250.8 | 125 | 67 | solver33 | 12395.0 | 136 | 0 | solver7 | 12173.6 | 0 | 66 |
| solver5 | 23888.5 | 143 | 83 | solver5 | 14952.3 | 143 | 0 | solver8 | 12669.9 | 0 | 67 |
| solver17 | 29790.5 | 153 | 96 | solver17 | 16959.1 | 153 | 0 | solver16 | 14427.5 | 0 | 95 |
| solver20 | 31482.8 | 111 | 66 | solver42 | 17869.9 | 130 | 0 | solver17 | 14525.7 | 0 | 96 |
| solver22 | 32660.0 | 167 | 100 | solver20 | 18487.5 | 111 | 0 | solver20 | 14637.9 | 0 | 66 |
| solver16 | 33840.8 | 156 | 95 | solver22 | 18887.6 | 167 | 0 | solver22 | 15385.3 | 0 | 100 |
| solver39 | 37601.4 | 169 | 97 | solver40 | 18975.0 | 159 | 0 | solver39 | 18849.0 | 0 | 97 |
| solver19 | 41213.5 | 157 | 105 | solver6 | 19257.6 | 158 | 0 | solver18 | 20056.9 | 0 | 105 |
| solver18 | 41719.3 | 158 | 105 | solver16 | 20078.4 | 156 | 0 | solver19 | 20216.5 | 0 | 105 |
| solver21 | 43555.7 | 167 | 109 | solver21 | 20919.6 | 167 | 0 | solver21 | 24387.0 | 0 | 109 |
| solver6 | 49476.0 | 158 | 113 | solver39 | 21040.8 | 169 | 0 | solver41 | 25253.1 | 0 | 111 |
| solver26 | 51536.4 | 163 | 136 | solver19 | 22894.1 | 157 | 0 | solver26 | 31366.9 | 0 | 136 |
| solver40 | 52064.5 | 159 | 119 | solver18 | 23502.3 | 158 | 0 | solver37 | 33065.6 | 0 | 129 |
| solver41 | 55428.7 | 182 | 111 | solver26 | 23913.2 | 163 | 0 | solver40 | 34597.6 | 0 | 117 |
| solver36 | 56951.3 | 20 | 78 | solver37 | 29366.8 | 195 | 0 | solver6 | 35227.7 | 0 | 113 |
| solver37 | 60869.3 | 195 | 130 | solver41 | 31656.8 | 182 | 0 | solver34 | 46427.0 | 0 | 145 |
| solver34 | 79069.8 | 173 | 145 | solver34 | 38063.1 | 173 | 0 | solver36 | 55211.4 | 0 | 78 |

1 Introduction

Over recent years, the importance of the international SAT competition has grown to being an awaited event in the community. The major impact of being ranked among the best solvers is beneficial both for academic and industrial competitors. As a consequence, the scoring scheme of the competition needed some more formal basis.

The method described in this paper is designed to overcome some of the drawbacks observed in earlier methods. The primary difficulty is that, because the underlying problem requires exponential time in practice, one must either set very easy problems to be sure all solvers can succeed, or one must allow for the fact that some solvers will not succeed on some instances. It is commonly agreed that the first alternative does not lead to interesting outcomes.

The paper outline is as follows. After presenting the design objectives and discussing drawbacks with current approaches, we describe the purse-based method that was decided upon. Some properties of this purse-based method are discussed. Then we take some examples from stage one of the SAT 2005 competition to illustrate how the scoring scheme works and how the rankings would change if alternative ranking schemes were used. Preliminary experimental results are presented for the first stage of the SAT 2005 Competition, involving about 30 solvers and hundreds of benchmark instances. The paper concludes with a brief discussion of the critical issues regarding the new scoring scheme and provides a first assessment on how it can be improved.

2 Design Objectives

One key idea behind the SAT competition is to award a solver that is good on a wide range of SAT instances. In the previous year of the competition, this was implemented using a scoring scheme that ranked the solvers with a tiered system: First, the solvers were ranked by being able to solve *some instance* in a highest number of different series. Ties were then broken using the total number of benchmarks solved. Unfortunately, in this system there is no difference between solving a benchmark solved by all solvers or one solved by only a few solvers. The same applies to series too.

Another key idea of the competition was to focus on solvers that are the *only* ones to solve some benchmarks: in the SAT and CASC competitions, those solvers are called *state-of-the-art contributors* (abbreviated SOTAC). In the previous scoring scheme, the solvers did not benefit directly for being SOTAC in their category, even though SOTAC solvers were usually among the top ranked solvers.

Third, the time needed to solve a given benchmark also needs to be considered. While the CPU time was indirectly used for scoring the solvers in the previous years of the SAT competitions, by using a fixed timeout per benchmark, there was no way to discriminate among the solvers able to solve a given benchmark within that timeout.

Furthermore, the second stage ranking was based only on the number of benchmarks solved during the second stage, among those benchmarks that had not been solved by *any solver* during the first stage. This criterion is based on very strong assumptions:

- The remaining benchmarks are representative of the initial set of benchmarks.
- The solvers will behave in the second stage in a way similar to the first stage.

However, these assumptions did not necessarily hold. Although it is likely that the winners of the previous competitions could have been declared winners using various scoring schemes, nevertheless, the rankings of the remaining top solvers could have changed a lot.

The scoring scheme used for the SAT 2005 competition is designed to address these issues. It incorporates these features:

- It gives more credit for solving hard benchmarks than solving easy ones.
- It gives more credit for solving a benchmark fast.
- It gives extra credit for each series solved.

Table 3: RANDOM, best performers last.

| (A) SAT+UNSAT | | | | (B) SAT | | | | (C) UNSAT | | | |
|---------------|---------|------------|-------|----------|--------|------------|-------|-----------|-------|------------|-------|
| Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | | Solver | Score | Nbr Solved | |
| | | Sat | Unsat | | | Sat | Unsat | | | Sat | Unsat |
| solver36 | 0 | 0 | 0 | solver36 | 0 | 0 | 0 | solver1 | 0 | 0 | 0 |
| solver5 | 349.5 | 3 | 0 | solver5 | 349.5 | 3 | 0 | solver15 | 0 | 0 | 0 |
| solver16 | 357.2 | 3 | 0 | solver16 | 357.2 | 3 | 0 | solver16 | 0 | 0 | 0 |
| solver24 | 432.0 | 3 | 0 | solver24 | 432.0 | 3 | 0 | solver17 | 0 | 0 | 0 |
| solver17 | 547.3 | 5 | 0 | solver17 | 547.3 | 5 | 0 | solver18 | 0 | 0 | 0 |
| solver21 | 673.7 | 7 | 0 | solver21 | 673.7 | 7 | 0 | solver19 | 0 | 0 | 0 |
| solver20 | 881.7 | 5 | 0 | solver20 | 881.7 | 5 | 0 | solver20 | 0 | 0 | 0 |
| solver22 | 1247.9 | 10 | 1 | solver22 | 1155.7 | 10 | 0 | solver21 | 0 | 0 | 0 |
| solver40 | 1325.5 | 12 | 0 | solver40 | 1325.5 | 12 | 0 | solver24 | 0 | 0 | 0 |
| solver37 | 1457.2 | 17 | 0 | solver37 | 1457.2 | 17 | 0 | solver27 | 0 | 0 | 0 |
| solver18 | 1891.7 | 17 | 0 | solver6 | 1621.0 | 20 | 0 | solver28 | 0 | 0 | 0 |
| solver39 | 2069.6 | 16 | 0 | solver18 | 1891.7 | 17 | 0 | solver31 | 0 | 0 | 0 |
| solver19 | 2182.8 | 19 | 0 | solver39 | 2069.6 | 16 | 0 | solver32 | 0 | 0 | 0 |
| solver6 | 2329.3 | 20 | 7 | solver19 | 2182.8 | 19 | 0 | solver36 | 0 | 0 | 0 |
| solver33 | 5540.5 | 35 | 15 | solver34 | 3251.8 | 33 | 0 | solver37 | 0 | 0 | 0 |
| solver34 | 6273.5 | 33 | 21 | solver33 | 3580.9 | 35 | 0 | solver39 | 0 | 0 | 0 |
| solver26 | 6859.1 | 34 | 22 | solver26 | 3658.6 | 34 | 0 | solver40 | 0 | 0 | 0 |
| solver38 | 9818.2 | 38 | 29 | solver38 | 3802.0 | 38 | 0 | solver42 | 0 | 0 | 0 |
| solver8 | 14590.1 | 50 | 34</ | | | | | | | | |