

# Search-Based Estimation of Problem Difficulty for Humans

Matej Guid, Ivan Bratko

Faculty of Computer and Information Science, University of Ljubljana, Slovenia

**Abstract.** The research question addressed in this paper is: Given a problem, can we automatically predict how difficult the problem will be to solve by humans? We focus our investigation on problems in which the difficulty arises from the combinatorial complexity of problems. We propose a measure of difficulty that is based on modeling the problem solving effort as search among alternatives and the relations among alternative solutions. In experiments in the chess domain, using data obtained from very strong human players, this measure was shown at a high level of statistical significance to be adequate as a genuine measure of difficulty for humans.

**Keywords:** human problem solving, heuristic search, problem difficulty

## 1 Introduction

In this paper, we address the research question: Given a problem, can we automatically predict how difficult the problem will be to solve by humans? This question is complex and concerns many aspects. It depends on the type of problem and on the human's knowledge about the problem domain. Our current investigation is focused on problems in which the difficulty arises from the combinatorial complexity of problems. We propose a measure of difficulty that is based on modeling the problem solving effort as search among alternatives and the relations among alternative solutions.

The basis for that is the AI formulation of problem solving as search: a given problem is reduced to finding a path in the state space. This typically leads to the problem of combinatorial complexity due to the rapidly growing number of alternatives. To overcome this problem, *heuristic search* is widely used. For the nodes in the state space heuristic estimates are determined, indicating how promising nodes are with respect to reaching a goal node, and this knowledge then guides the search.

Our experiments in this paper with the proposed measures of difficulty were carried out in a game playing domain (chess). Our method is based on heuristic search. In general, relatively little research has been devoted to the issue of problem difficulty. Some specific puzzles were investigated with this respect, including Tower of Hanoi [1], Chinese rings [2], 15-puzzle [3], Traveling Salesperson Problem [4], Sokoban puzzle [5], and Sudoku [6]. To the best of our knowledge, no related work deals with possibilities of using heuristic-search based methods for determining how difficult the problem is for a human.

## 2 Method

Our basic idea is as follows: a given problem is difficult with respect to the task of accurate evaluation and finding the best solution, when different “solutions,” which considerably alter the evaluation of the initial problem state, are discovered at different search depths. In such a situation a human has to analyze more continuations and search to a greater depth from the initial state to find actions that may greatly influence the assessment of the initial state, and then eventually choose the best continuation [7].

In the experiments, the chess program HOUDINI 1.5a (64-bit), one of the strongest chess engines, was used to analyze more than 40.000 positions from real games played in World Chess Championship matches, using the methodology presented in [8]. Each position was searched to a fixed depth ranging from 2 to 20 plies. The aim of the heuristic search performed by the engine was both (I) to obtain the data for experimental evaluation of our proposed difficulty measure called “difficulty score,” and (II) to estimate players’ errors in these positions. A large data set made it possible to obtain average players’ deviations from best play across a wide range of positions with the same difficulty score.

### 2.1 Proposed Measure of Difficulty


In accordance with our hypothesis about what makes the problems difficult for a human, an algorithm for calculating the difficulty of a chess position had to satisfy the following properties:

1. A problem is difficult if several different sensible “solutions” appear with increasing depth of search. That is, different amounts of search produce different solutions of the problem.
2. The higher the magnitude of differences in the values of various “solutions” obtained at different search depths, the greater the difficulty of the problem.

A formal measure of difficulty that attempts to implement the principles above is given by the following formula.

$$\sum_{d=3}^{MAX} |E(best_d) - E(second\_best_d)| \times [best_d \neq best_{d-1}] \quad (1)$$

where  $best_d$  is the move that the chess program suggests as best at  $d$ -ply search,  $E(best_d)$  and  $E(second\_best_d)$  are the evaluations of the best and the second best move (respectively) at depth  $d$ , and  $MAX$  is a user-defined parameters for the maximal search depth used by the program. The bracket value  $[\ ]$  is 1 if the condition holds, otherwise it is 0. We call this measure the difficulty score. Figure 1 illustrates how the difficulty score is calculated.



d	best	E1	second	E2	DS
2	Nf3-g5	123	Qd1-c2	80	–
3	Nf3-g5	107	Qd1-c2	103	0
4	Nf3-g5	117	Qd1-c2	103	0
5	Nf3-g5	117	Qd1-c2	103	0
6	Nf3-g5	117	Qd1-c2	103	0
7	Nf3-g5	117	Qd1-c2	103	0
8	Nf3-g5	98	Qd1-c2	98	0
<b>9</b>	<b>Qd1-c1</b>	<b>118</b>	<b>Nf3-g5</b>	<b>92</b>	<b>26</b>
10	Qd1-c1	163	Qd1-c2	128	26
11	Qd1-c1	178	Qd1-c2	166	26
<b>12</b>	<b>Qd1-d4</b>	<b>805</b>	<b>Qd1-c2</b>	<b>166</b>	<b>665</b>

**Fig. 1.** Euwe-Alekhine, 16<sup>th</sup> World Chess Championship, Game 14, position after Black's 19<sup>th</sup> move. The table on the right shows the values of  $best_d$ ,  $E(best_d)$ ,  $second\_best_d$ ,  $E(second\_best_d)$ , and the difficulty score, respectively, for each search depth  $d$  in range from 2 to 12 plies. At  $MAX = 12$ , formula (1) thus assigns this position the difficulty score of 665. In the game Euwe, the contender for the title of World Champion, failed to find the strongest move 20.Qd1-d4, with a winning attack.

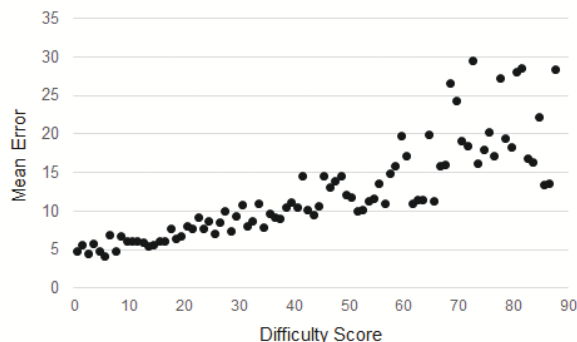
### 3 Results

To evaluate the adequacy of our proposed measure of difficulty, we carried out the following experimental evaluation. If the difficulty score indeed measures the difficulty of a chess position for human chess players, then a high difficulty score of a given position should indicate a relatively high probability of a human player making a mistake in that position. Also, a higher difficulty score should indicate a more severe error. This was experimentally tested by observing the correlation between the difficulty scores of positions and the error scores of very strong chess players in these positions. As mistakes by very strong players are subject to chance it was appropriate to average the errors in *sets* of positions with similar difficulty scores.

Figure 2 shows the relation between the difficulty scores (that is the predicted difficulties of chess positions), and the players' mean errors in positions with (roughly) the same difficulty score. Ideally, the mean error should be a monotonically increasing function of difficulty score. Because of the randomness of human errors, this relation has to be tested statistically. A Spearman's correlation was run to determine the relationship between the difficulty scores and the mean errors. There was a very strong, positive monotonic correlation between Difficulty Score and Mean Error ( $r = .93$ ,  $n = 88$ ,  $p < .001$ ).

### 4 Conclusions

Our approach to predicting the difficulty of problems for humans is based on modeling the problem solving as search. We proposed a concrete measure of difficulty, called difficulty score. It was experimentally shown to be statistically



**Fig. 2.** The scatter plot above shows the relation between the predicted difficulty (obtained with formula (1),  $MAX = 15$ ) and mean players' error in chess positions with corresponding difficulty scores. Each data point is represented by at least 30 examples.

adequate as a genuine measure of difficulty for humans. The experiments were carried out in the domain of chess using the experimental data obtained from extremely strong human experts - world chess champions. It should be noted that despite high overall statistical significance of the proposed measure, the success of difficulty score as a reliable predictor of the difficulty of individual problems is open to further investigation. This will probably depend on the application. Also, the implementation by a concrete difficulty measure of the two basic assumptions about the measures' properties is open to refinements. For example, it might be better (I) to consider that decision changes become more and more important with increasing search depth, and (II) to take into account more than just two best solutions as it is done in formula (1).

## References

1. Kotovsky, K., Hayes, J., Simon, H.: Why are some problems hard? Evidence from tower of Hanoi. *Cognitive Psychology* **17**(2) (1985) 248–294
2. Kenneth Kotovsky, H.A.S.: What makes some problems really hard: Explorations in the problem space of difficulty. *Cognitive Psychology* **22**(2) (1990) 143–183
3. Pizlo, Z., Li, Z.: Solving combinatorial problems: The 15-puzzle. *Memory and Cognition* **33**(6) (2005) 1069–1084
4. Dry, M., Lee, M., Vickers, D., Hughes, P.: Human performance on visually presented traveling salesperson problems with varying numbers of nodes. *Journal of Problem Solving* **1**(1) (2006) 20–32
5. Jarušek, P., Pelánek, R.: Difficulty rating of sokoban puzzle. In: Proc. of the Fifth Starting AI Researchers' Symposium (STAIRS 2010), IOS Press (2010) 140–150
6. Pelánek, R.: Difficulty rating of sudoku puzzles by a computational model. In: Proc. of Florida Artificial Intelligence Research Society Conference (FLAIRS 2011), AAAI Press (2011) 434–439
7. Guid, M., Bratko, I.: Computer analysis of world chess champions. *ICGA Journal* **29**(2) (2006) 65–73
8. Guid, M., Bratko, I.: Using heuristic-search based engines for estimating human skill at chess. *ICGA Journal* **34**(2) (2011) 71–81