NOTE

DATA-MINING CHESS DATABASES

E. Bleicher, G.M^{c}C. Haworth[1] and H.M.J.F. van der Heijden

Berlin, Germany; Reading, U.K., and Deventer, The Netherlands

ABSTRACT

This is a report on the data-mining of two chess databases, the objective being to compare their sub-7-man content with perfect play as documented in Nalimov endgame tables. Van der Heijden's ENDGAME STUDY DATABASE IV is a definitive collection of 76,132 studies in which White should have an essentially unique route to the stipulated goal. Chessbase's BIG DATABASE 2010 holds some 4.5 million games. Insight gained into both database content and data-mining has led to some delightful surprises and created a further agenda.

1. INTRODUCTION

Harold van der Heijden's (2010) ENDGAME STUDY DATABASE IV contains 76,132 studies in which White is challenged either to draw or to win. Chess studies should conform to some basic criteria, the main one being that the solution, the successful line, exists and is essentially unique. Given the Nalimov DTM[2] EGTs with their perfect information about sub-7-man (s7m) chess, the following questions arise:

- do the studies' mainline s7m-positions have values compatible with the stipulated goals?
  clearly, all such positions in 'draw (win) studies' should be draws (wins) for White,
- does White play DTM-optimal moves in win studies, and how significant is it if not?
- if there is more than one DTM-optimal move, are they essentially the same or not?

The CQL query language (Costeff and Stiller, 2003) had in fact been created to identify studies in Harold's corpus with certain chessic properties. It can trivially identify all studies containing sub-*m*-man play and two subsets of studies with draw (win) stipulations and s7m-play in the mainline were duly created.

PGN2FEN (Foden, 2010) converts pgn-format files into lists of FEN positions but does not identify the position-subsets corresponding to each study sequence. This lacuna was fixed using only EXCEL spreadsheet functionality by identifying where position details (move-#, stm, number of men and pawns) indicated the start of a new study. Positions with more than six men were useful for this and were then removed.

Chessbase's BIG DATABASE (2010) was treated in the same way[3]. Here, the questions are about fallible play:

- how many games reach s7m-chess with value 1-0, = or 0-1 end with value 1-0, = or 0-1?
- what is the pattern of concession of theoretical value in s7m-play?
- what is the DTM-profile of the DTM-deepest positions?
- what is the frequency of zugzwang positions of Type A and B-DTM[4]?

This second exercise underlined a basic data-mining principle: use the right tools to store and process your data. The scale of the computation rather overpowered EXCEL technology with its limit of ~1m rows/worksheet. A combination of greater computing literacy, GAWK and a relational database would for example have been a more strategic and efficient approach than 50^+ worksheets. However, the derived sets of positions were eventually sent to Eiko Bleicher who provided, in the EGT-analysis of each position, the theoretical value, DTM, and the number of (optimal, winning, drawing, losing) moves, both *stm* and *sntm*.

[1] The University of Reading, Berkshire, UK, RG6 6AH. email: guy.haworth@bnc.oxon.org.

[2] DTM ≡ Depth to Mate, a metric in which depth is measured in winner's moves in DTM-minimaxing play.

[3] Our thanks to Ken Regan who carried out the initial CQL subsetting of BIG DATABASE 2010.

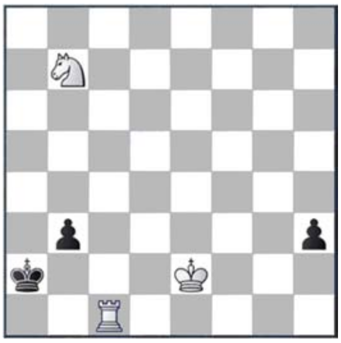
[4] In Type A zugzwang positions, the side-to-move gains value by playing a null move; in Type B-DT*x* zugzwangs, the stm merely wins more quickly or loses more slowly in terms of metric DT*x*. Here, DT*x* ≡ DTM.

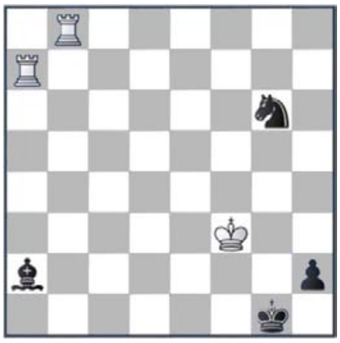
2. RESULTS FROM ENDGAME STUDY DATABASE IV

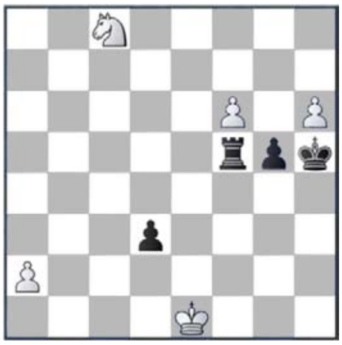
The data-mining exercise was first run on ENDGAME STUDY DATABASE III and then on the content anticipated for ‘IV’. Harold was able to improve edition IV in a remarkably short time, converting the EGT-findings into discoveries about mistranscriptions (the data curator’s burden) and into over 5,100 notes about better moves generously ascribed to ‘GH/EB’. 45,392 (59.62%) of all studies reach s7m-play. Of these, 3,068 (6.76%) were faulted by EGT-analysis, about half of them not previously cooked. 507,091 mainline positions were evaluated of which 18,741 (3.70%) had values incompatible with their study’s stipulation.

The Nalimov EGTs do not include positions with castling rights and Bleicher’s automated analysis ignores such rights. Manual analysis of positions with castling rights was therefore necessary, especially as in studies these rights tend to be essential to the plot and cause studies to be incorrectly EGT-faulted. There is even one previously unnoted game in BIG DATABASE 2010 with s7m-castling rights[5] so the inclusion of positions with castling rights in EGTs seems overdue. Since one cannot move from a position to one with more castling rights, the small ‘castling rights’ EGTs can be computed without recomputing current EGTs.

The analysis of the data-mined results revealed many intriguing ‘escapes’ for Black, some of the best discussed in van der Heijden et al. (2010). Harold provides three more interesting, busted win studies below. Bleicher (2010) is of course useful here to confirm the value of and options in the s7m-positions.


db#52,221: Y. Hoch (1982)
2nd comm., *Schakend Nederland*


db#22,772: V. Villani (1947)
L'Italia scacchistica


db#5,119: G. Petenyi (1908)
Az Ujsag

Hoch intended the solution **1.Rc4 b2 2.Ra4+ Kb3 3.Nc5+ Kc3** followed by **4.Ra3+ Kc2 5.Na4! b1Q 6.Rc3** mate. However, the study-like **4.Ne4+!** wins even more quickly: **4...Kb3 5.Rd4** and now, with the bK at b3, the promotions **b1Q/R/B** fail to the forking check **Nd2+**: **4…b1N** also fails to **5.Kd3 h2 6.Nc5+**. Villani intended the solution **1.Rh8 Nxh8 2.Rxa2** and now the presence of bNh8 seems to prevent 2...h1Q+ 3.Kg3 and Black is unable to cover a1. Therefore the composer proposed **2...h1N 3.Rg2+ Kf1 4.Rh2 Kg1 5.Rxh8 Nf2 6.Ra8 Nd3 7.Ke3** wins. But 2...h1Q+ 3.Kg3 and now 3...Kf1 4.Ra1+ Ke2 5.Rxh1 Nf7 draws. This means that 2...h1R and 2...h1B also draw, i.e., every promotion draws except the one the composer used! Lastly, Petenyi intended **1.Ne7! Rxf6 2.h7 Rh6 3.Ng8! Rxh7 3.Nf6+ Kg6 5.Nxh7** followed by **5...Kxh7 6.Kd2! Kg6 7.a4 g4 8.Kxd3 Kf5 9.a5** winning. However, Black has the win: 5...g4! 6.Nf8+ Kf5 7.Kd2 (or 7.Nd7) 7...g3!

Bleicher’s EGT-analysis highlighted all DTM-suboptimal play by White in win studies, alternative DTM-suboptimal winning moves, multiple DTM-optimal moves and Type A/B zugzwangs. Our mathematical and chessic analysis of this data continues and could usefully involve the studies community via the web on a more inclusive basis. DTM-suboptimal play by White in win studies does not always indicate a significant cook. However, intuition suggests that the shallower the DTM of a position and the greater the DTM-concession at that point, the more likely it is that there is indeed a fault in the study. There are in the corpus 25, 67, 90, 129, and 172 missed mates in one to five respectively.

Alternative DTM-suboptimal winning moves might lead to significantly different winning lines – or might merely be *time wasting* moves, i.e. allow Black to force White, ambitious to achieve the goal, to return to the position just left. The latter are held to be less serious, even minor, duals. Haworth (2009) defined an algorithm for determining whether alternative moves are time-wasting moves or not. This is based on variants of chess in

[5] Lebeda-Erben, Wuertenberg (2005). Admittedly an under-10 game but an *otb fact* nevertheless. White mated, castling at 8/8/8/8/4Q2P/5R2/1k2K2R w K - 0 56: the play suggests this had been the intended goal for some time.

which the position in question is deemed to be worth a ½-point less to White than it actually is. A variant-EGT is then (partially) created to resolve the question of moves' time-wasting status. This is not a major task today given the existence of the DTM EGTs, bit-based EGTs and multicore computers.

The investigation of zugzwangs is now part of a wider s7m-zugzwang study following the definition of Type A, B and C[6] zugzwangs (Bleicher and Haworth, 2010). There are 22,882 s7m-zugzwangs in the database: 7,690 Type A (293 *full-point* Type A3) and 15,192 Type B-DTM.

3. RESULTS FROM BIG DATABASE 2010

This database comprises some 4.5 million games dating from 1560 to 2009, again with the inevitable sprinkling of errors and omissions caveating the results below. The searchlight of perfect knowledge is unforgiving but reveals remarkable passages of accurate play from notionally fallible, top level players. 266,236 (5.97%) of these games reach the s7m-zone. Of these, 232,967 (87.50%) involved no concession of theoretical value at all, the result being largely determined but not agreed before the s7m-zone. A further 8,771 games (3.29%) ended with no net change in theoretical value in the s7m-zone though both sides gave and spurned chances. 22,332 games (8.39%) saw a ½-point net change and 404 games (0.15%) a 1-point net change in theoretical value during the s7m-play. Remaining games entered the s7m-zone with a lone King.

Some clear mistranscriptions, including apparently inexplicable moves, game 'results' and instances of men repeatedly left en prise, were revealed and communicated to Chessbase. The sharp-eyed Ken Regan also sent us an example[7] of the post-1997 influence of DGT technology on game-scores. Its associated convention of *centring Kings* to indicate an agreed result can be mistaken for a move. Last moves centring Kings after 1997 should perhaps be ignored. Overall, 7,339 games (2.76%) are given a result other than the theoretical value of the final position. They are worth classifying under the headings mistranscription, unnecessary resignation and generously agreed draw – a suitable challenge for a web-based chess community.

To close, here are some of the remarkable outposts found. Of 4.22m s7m-positions, only 75 have DTM > 150:[8] endgames KRBKNN and KRRPKQ are to the fore. At the other extreme, 1,914 mates in one were missed. In KNNKP, concessions of 925 units of DTM-depth failed to change a 1-0 result. KPPPKR endgames saw the win transferred from one side to the other in more than one game. Several games see players between them conceding more than 10 points of theoretical value. After further analysis, data-visualisation and web-publication, there will be much for the chess community to wonder at, comment on, and ponder.

References

Bleicher, E. (2010). Sub-7-man chess EGT-query service.

Bleicher, E. and Haworth, G.M^{c}C. (2010). *6-man chess and zugzwangs.* Proceedings of *Advances in Computer Games, 12* (2009). Lecture Notes in Computer Science, 6048, pp. 123-135. ISSN 0302-9743.

Chessbase (2010). http://www.chessbase.com/shop/product.asp?pid=475. BIG DATABASE 2010.

Costeff, G. and Stiller, L. (2003). http://www.rbnn.com/cql/ CQL: Chess Query Language.

Foden, T. (2010). http://www.pgn2fen.com-about.com/ The PGN2FEN v1.0.4 format-conversion utility.

Haworth, G.M^{c}C. (2009). *The Scorched Earth algorithm.* A presentation to the Chess Endgame Study Circle, Pushkin House, London, UK, 2009-04-03.

van der Heijden, H. (2010). http://www.hhdbiv.nl/. ENDGAME STUDY DATABASE IV.

van der Heijden, H., Bleicher, E. and Haworth, G.M^{c}C. (2010). *Endgame Table Analysis of Chess Studies I-II,* EG Vol. 16, No. 180, pp. 114-119 and No. 181, pp. 163-169.

[6] In Type C zugzwangs, a *null move* is arguably (or given *model fallible players*, provably) the best move.

[7] Onischuk-Rodshtein (Olympiad 2010), 8/8/7p/1P6/3kp3/7P/4K3/8 b 66: the latter apparently plays the losing 66. … Ke5?? rather than the comparatively obvious drawing Kc5 or even Kd5 - before agreeing a draw.

[8] Deepest is KRBKNN 8/8/6k1/r7/3NNK2/8/8/3b4 w 54 (DTM = 209): Karjakin-Shirov (2007), drawn in KRKN.