Adapting the ELO rating system to competing subpopulations in a "man-hill"

Grégory Valigiani [a,b] Evelyne Lutton [b] Pierre Collet [a]
[a] *LIL Lab - ULCO - 62228 Calais - France*
[b] *COMPLEX Team - INRIA Rocquencourt - 78150 Le Chesnay - France*

Abstract. Paraschool (the French leading e-learning company, with more than 250,000 registered students), wanted an intelligent software to guide students in their graph of pedagogic items. The very large number of students suggested to use students as artificial ants, leaving stigmergic information on the web-site graph to optimise pedagogical paths. The differences between artificial ants and students led to describe a new concurrent paradigm called "man-hill optimization," where optimization emerges from the behaviour of humans exploring a web site.

At this stage, the need of rating pedagogical items showed up in order to direct students towards items adapted to their level. A solution was found in the ELO [12] automatic rating process, that also provides (as a side-effect) a powerful audit system that can track syntactic and semantic problems in exercises. For an effective use, this paper shows how the ELO rating process has been modified to overcome the Deflation problem.

Keywords. E-Learning, Ant Colony Optimization, "Man-Hill" Optimization, concurrent optimization, ELO Rating, Turnover, Sub-pools.

Introduction

Paraschool is currently the French leading e-learning company, with more than 250,000 registered students. Back in 2002, Paraschool was looking for a system that could enhance web-site navigation by making it intelligent and adaptive to the user. Since the tree of available excercises could be turned into a graph visited by students (where pedagogical items are nodes and hypertext links are arcs), Ant Colony Optimization (ACO) techniques (a concurrent optimization paradigm [4,1,2]) could apply and show interesting properties: adaptability and robustness.

Unfortunately, real-size experimentations have shown that ant-hill optimization techniques developed in Paraschool do not directly apply because students do not behave like artificial ants. The concept of an artificial "student-hill," or more generally "man-hill," has been introduced and analysed [7,8,9].

In a refinement stage[10], the level of items and students needs to be evaluated in order to direct students towards exercises of matching level (there is no point in suggesting an exercise that is overly difficult or simple to a particular student). The Paraschool pedagogical team could rate the different items based on their

knowledge and experience, but what may seem simple for a teacher may seem difficult for a student. Moreover the level of the students must also be evaluated, which is quite difficult if the student does not have a long enough interaction with a human teacher.

A solution to this very important problem was found in the chess world, with the automatic ELO rating computation. After a short description of the Paraschool "man-hill" concurrent optimizer, the chess ELO rating is described in section 2 and then applied to Paraschool system in section 3. Results over 4 years of data show that the ELO evaluation process can be modified to overcome the known problems of the ELO system, thanks to the specificities of the e-learning system.

1. The Paraschool "man-hill"

1.1. Ant Colony Optimization

The Paraschool e-learning software is used in French schools or by individual students at home over the Internet. Connected students have access to thousands of pedagogic items (know-hows, lessons, drills) that were originally deterministically related by hypertext links.

The aim of the presented work is twofold:

1. find the best succession of items to maximize learning, and
2. insert some intelligence into the system so that different students have a different view of the Paraschool software.

ACO (developed after the observation of ant-hills [6,3]) uses virtual ants to find minimal paths in a graph. In the Paraschool system, the very large number of students triggered the idea to apply a similar technique using real students rather than virtual ants, with the aim of optimizing pedagogical paths traversing a set of educational topics. Students release artificial pheromones on the graph, depending on how they validated an item (success or failure). This stigmergic information can then be used by other students to choose their way on the different possible pedagogical paths.

Developing an ant colony optimization technique using human students on the Paraschool graph has however led to the (obvious) conclusion that humans do not behave as natural or artificial ants:

- There is no control on human students as on artificial ants.
- Artificial ants are permanently active on the entire environment, to the contrary of students (holidays, navigation per topics along the year).
- Social insects are inherently altruistic, while human users are individual by essence: they do not like to be treated identically, and on the contrary, appreciate systems that are adapted to their particular case.

Tests have shown that because of these differences, the standard ACO paradigm does not work straight out of the box. The concept of "man-hill" optimization has therefore been introduced. Problematic pheromone evaporation dur-

ing periods of inactivity over some areas of the graph has been solved by a new concept of pheromone *erosion*, and the need for individuality is dealt with thanks to the introduction of *multiplicative pheromones*, that only belong to a particular student. A further refinement allowing to tailor the system for a specific student is to take into account the level of the student, and direct him toward exercises he has a reasonable chance to solve. In order to achieve this, one must find a way to rate the drills and the students.

2. Using an ELO rating scheme in an interactive tutoring system

One could think of several ways to rate the respective difficulty of a drill and the proficiency of a student. The first idea that comes to mind is to ask the teachers who wrote the items to rate them on a scale going from easy to difficult. An experiment over 45 items has been done with two different teachers who were asked to evaluate items on a scale from 1 to 6. It appears that 8 evaluations did not reflect the real success rate of students on the item and 16 other evaluations were not quite right. This method tends to be error-prone because it relies on the judgment of the teacher, and on the level of the student that is faced with the drill. A much better system would be an automatic rating process for both items and students, but such a thing is very difficult to calibrate. The chosen solution was to use a very refined system called the ELO rating [12], that has been used in the Chess community for the last 50 years, where individuals compete against each other on a regular basis. At the end of the fifties, a mathematician, A. E. ELO [12], developed a chess rating system, based on the Thursone Case V Model [11] which has been adopted by chess federations worldwide. The ELO system was successful, due to the fact that rating differences between two competitors ($s_i - s_j$) and mutual winning chances are much more clearly related in this system than in any other. Moreover, ELO was the first to use computers for his calculations, which enabled him to rate a huge amount of players.

2.1. Rating update

The equation $S_i(t+1) = S_i(t) + K(R_{ij} - R_{ij_e})$ describes how an original rating $S_i(t)$ is updated as a function of the expected outcome R_{ij_e}. If i and j are rated players, one can logically expect the stronger to win over the weaker. The expected outcome is called R_{ij_e}. However, the real outcome of the game R_{ij} may be different.

If $R_{ij} = R_{ij_e}$, the rating of the players was accurate. If $R_{ij} \neq R_{ij_e}$, the ratings $S_i(t)$ and $S_j(t)$ need to be updated to reflect the outcome of the game.

The impact of the $R_{ij} - R_{ij_e}$ difference is tuned thanks to a variable K, which represents the maximum amount of points that can be won in one game. A high K-factor gives more weight to new results while a low value increases the influence of earlier performances. The K-factor fluctuates between 16 for great players (ELO-rate$>$ 2400) and 32 for weak ones (ELO-rate$<$ 2100).

According to the Bradley-Terry Model[11], if the rating difference ($S_i(t) - S_j(t)$) is known between players i and j, the expected probability of success of player i against player j can be defined as:

$$R_{ij_e} = \frac{1}{1 + 10^{\frac{S_i(t) - S_j(t)}{400}}}$$

This is the basic formula for the rating system of the United States Chess Federation.

In the Paraschool system, one can consider that students and exercises "compete" against each other, with the nice outcome that one can objectively compute their respective ELO rating, independently of any biases.

2.2. Inflation and Deflation

Since the introduction of the ELO rating system in the world of Chess, some problems arose because of:

Turnover : If no individuals enter or leave the pool of rated players, then every gain in rating by one player would (ideally) result in a decrease in rating by another player by equal amount. Thus, rating points would be conserved, and the average rating of all players would remain constant over time. But, typically, players who enter the rating pool are weaker than players who leave it. The net effect is this flow of players lowers the overall average rating.

Sub-pools : Inflation and deflation does not only occur in the rating pool as a whole but also within subpools. A subpool is a subset of players who keep playing together over longer periods of time without much contact with players outside their group. This results in subpools with artificially low or high ratings. Within the subpool, ratings may still have a reasonable predictive value, but as soon as players from a subpool enter larger tournaments, they will start winning/loosing many points quickly, until their ELO rating is readjusted with reference to the larger pool. Altogether, the subpool-phenomenon shows that it is important for players to periodically play against people outside of their sub-pool.

These factors question the "integrity" of the ELO system, as they can create a general inflation or deflation of the global ratings. The integrity of the system indicates to which extent a given rating s_i reflects a same level over time, and across different sub-pools.

3. ELO ratings inside the Paraschool System

Since the algorithm already works quite well in the chess environment, the same equations and parameters were used for Paraschool. As soon as a student rating has stabilized, applications are numerous:

1. Students have a way to know their level, and can visualize their evolution.
2. The Paraschool pedagogical team does not need to put a subjective artificial rating on each item.

3. A very interesting side effect is that the ELO rating can tell if a drill contains a semantic or pedagogic flaw (something very difficult to detect otherwise, when there are thousands of different items): if an item has an extremely high ELO rating, this shows that either there is an error in the exercise, making it impossible for students to solve it, or that the exercise is much too difficult for the students to solve (indicating a pedagogic flaw). The same goes for items with very low ELO values, that are either too simple for the students, or that can be solved using a bypass (not requiring the mental process planned by the teacher). The ELO rating of items revealed to be an invaluable aid to the Paraschool pedagogical team if considered as an audit system.
4. Finally (and that was the primary goal of the implementation of the ELO rating), the man-hill system can be refined to propose items adapted to the strength of a particular student.

3.1. Paraschool subpools

In Chess tournaments, any player can possibly compete with any other player, even though most competitions are held within specific countries.

In the Paraschool system, it is much less so for several reasons:

1. An item cannot compete against another item, and a student cannot compete against another student. This *de facto* creates two subpools, but of a different kind, where players can play exclusively with an individual of the other group. This peculiar dynamics is different from what occurs in the chess environment, and it can be used to find a way to get around the deflation problem (cf. below).
2. The Paraschool system also exhibits chess-like subpools, since it hosts students of different grades. After analysis, 95% of students in a grade exclusively compete with items of their grade. This leads to the conclusion that the student ELO rate may be inconsistent if a student tries to solve a problem of another grade. In this case, the decision was simply not to take into account a "match" between a student and an item of different grades. This means that 5% of information is lost, but the impact on the system is minor.

3.2. Turnover in Paraschool

As in Chess, turnover in Paraschool represents students entering or leaving the ELO rating system. These cases happen more often in the beginning/end of the school year. Normally a student should keep his account for several years. In practice, however, schools unfortunately update student lists and accounts every year, leading to possible turnover concerns.

On Fig. 1, the number of visits clearly shows periods of inactivity during summer vacations. In between, the average ELO rate of students tends to increase, which is a positive result (students are getting better). The drop in the beginning of each year comes from the fact that Paraschool increased its number of students

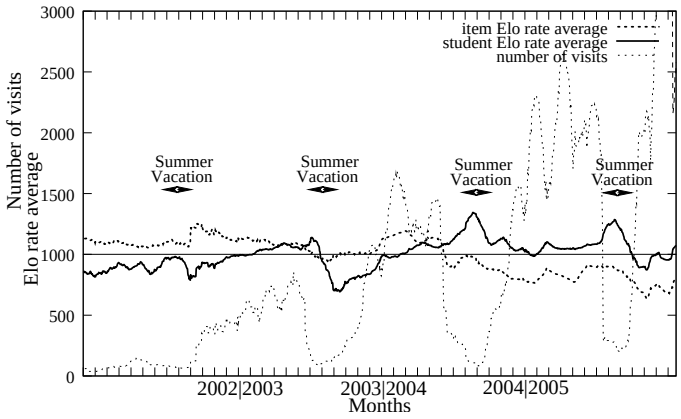

Figure 1. Average ELO Ratings and number of visits over a four years period.

from 50 000 to 250 000 over the four years on which data was collected (as can be seen by the increasing number of visits).

Fig. 1 also shows that the ELO rate of items tends to decrease year after year. This is because, to the contrary of schools (that reset student accounts every year), Paraschool does not reset the ELO rating of items, therefore causing a constant deflation of items ratings, as students get better over the years and steal ELO points to the items.

As seen above, the dynamics is different in the Paraschool system, since the system is dealing with two groups (the students and the items) that exclusively compete against each other. If one group has a stable ELO rating, this should stabilise the rating of the other group too.

The idea is then to apply different ratings for each group, in order to obtain greater stability and fight against natural deflation. For the students, the classical ELO rate system is kept. For the items, two options were studied:

Freezing: After a period of stabilization, the item gets its optimal rating and then deflation occurs. The goal is therefore to freeze the item before deflation begins. This means that once an item has its mature ELO rating, it keeps it forever, therefore stabilizing student ELO ratings at the same time. But the ELO rating of an item should also be computed from stabilized students.

Since the average number of visits per student (resp. item) is around 26 (resp 236), it was decided that student (resp. item) "maturity" would be obtained after 10 (resp. 75) evaluations.

On Fig. 2, the overall ELO gain is displayed depending on these two parameters: the number of evaluations after which an item is considered to have its optimal rating (*item_maturity*), and the number of evaluations after which a student is considered to have his optimal rating (*student_maturity*).

Probability-based ELO Rating for Items (PERI) If the classical ELO system can be seen as being too adaptive (therefore leading to deflation), the *freezing* method can be seen as being too static.

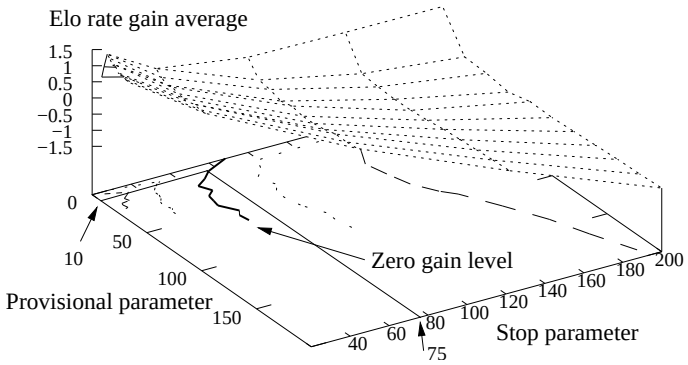

Figure 2. Average ELO gain over 4 years.

The ELO system is based on the fact that it is possible to compute a winning probability from the ELO ratings of two players. This means that if the winning probability of a player is known, one can evaluate his ELO rating by inverting the equation.
The PERI method computes a rating for items according to the success/failure ratio of students who tried to solve the item up to now. This means that the PERI rating is not subject to deflation, while at the same time, staying adaptive.

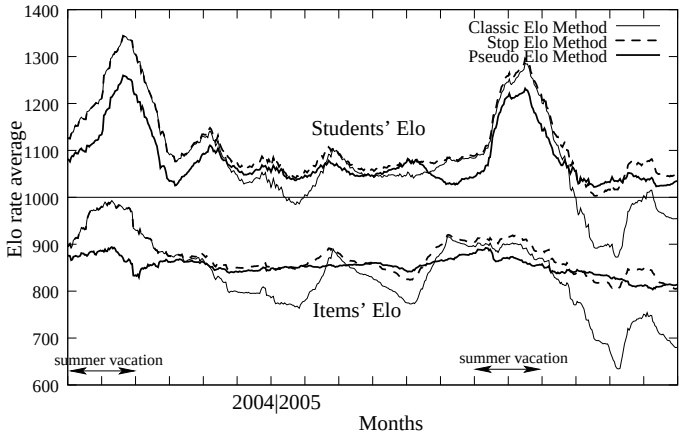

Figure 3. Average ELO Ratings for three ELO rating methods over a one year period.

Fig.3 shows the ELO rating of items and students during the previous school year (2004/2005) where the Paraschoom "man-hill" system should have found its stability. The thin solid curve represents the oscillating ELO rate of the standard method. The dashed curve shows the result of the *Freezing* method, while the bold curve represents the result of the PERI method. It appears that PERI is the most stable method over the year (with the *Freezing* method second, and the standard ELO rating third). It also appears that the students ELO rating (the

three upper curves) are oscillating with the three methods, but the range of the oscillations is smaller with the PERI method.

4. Conclusion

This paper introduces a new application for the ELO rating system, along with a new scheme allowing to prevent deflation of ELO points in the global system, as in the Paraschool concurrent "man-hill" system, it was possible to take advantage from the fact that one subpool was static (the items) to also stabilise the other subpool. By adding this automatic rating system, Paraschool hopes to get a good idea of the level of students and items, without the need for teacher evaluation.

This information can also be given to students, with two outcomes:

1. The students will get an idea of their proficiency, and will be able to follow their relative progression while using the system.
2. The E-learning system could observe the behaviour of each student when the possibility of choosing between a simple or a difficult item arises, and therefore get some indications on the pugnacity of the student, to even more specialise items suggestions (giving harder exercises to students who like difficulty, for instance).

References

[1] E.Bonabeau, M.Dorigo, G.Theraulaz, *Swarm Intelligence : From natural to Artificial systems*, Oxford University Press 1999, ISBN 0-19-513159-2.

[2] E. Bonabeau, M. Dorigo and G. Theraulaz, *Inspiration for optimization from social insect behaviour*, in Nature, vol. 406 pp 39–42, 2000.

[3] J.L. Deneubourg, S. Aron, S. Goss and J.M. Pasteels, *The self-organizing exploratory pattern of the argentine ant*, in Journal of Insect Behaviour, vol. 3 pp 159–168, 1990.

[4] M. Dorigo and G. Di Caro, *The ant colony optimization metaheuristic*, in New ideas in optimization, D. Corne, M. Dorigo and F. Glover (Eds), McGraw-Hill, pp 11–32, 1997.

[5] M. Dorigo, *Optimization, learning and natural algoritmhs*, PhD Thesis, politecnico di Milano, 1992.

[6] J.L. Deneubourg, J.M. Pasteels and J.C. Verhaeghe, *Probalistic behaviour in ants : a strategy of errors?*, in Theoritical Biology, vol. 105 pp 259–271, 1983.

[7] Y. Semet, Y. Jamont, R. Biojout, E. Lutton and P. Collet, *Artificial Ant Colony and E-Learning : An optimization of pedagogical paths*, HCI 2003.

[8] Y. Semet, E. Lutton and P. Collet, *Ant Colony Optimisation for E-Learning : Observing the emergence of pedagogic suggestions*, SIS 2003.

[9] G. Valigiani, Y. Jamont, R. Biojout, E. Lutton, P. Collet, *Experimenting with a Real-Size Man-Hill to Optimise Pedagogical Paths*, H. Haddad et al. Eds., Symposium on Applied Computing (SAC), ACM, Santa Fe, New Mexico.

[10] G. Valigiani, E. Lutton, Y. Jamont, R. Biojout, P. Collet, *Automatic Rating Process to Audit a Man-Hill*, WSEAS Trans. Advances in Engineering Education, 3, 1-7, 2006.

[11] R. A. Bradley, M. E. Terry, *The Rank Analysis of Incomplete Block Designs*, The Method of Paired Comparisons, Biometrika, 39, 324-345, 1952.

[12] A. E. Elo, *The rating of chess players past and present*, New York: Arco Publishing, 1978.