An Intelligent Search Technique for Solving Train Scheduling Problems: Simulated Annealing and Constraint Satisfaction

M.T. Isaai

This paper presents a hybrid scheduling technique for generating the predictive schedules of passenger trains. The algorithm, which represents a combination of simulated annealing and a constraint-based heuristic, has been designed using an object-oriented methodology and is suitable for a primarily single-track railway with some double-track sections. The search process gets started from a good initial solution created by the scheduling heuristic and continues, according to the simulated annealing search control strategy. The heuristic is also used in the neighborhood exploration process. This hybrid approach solves the problem in a short span of time. Simulation experiments, with the real data of manual timetables and two corridors of Iran’s railway, show the superiority of the hybrid method to the heuristic designed and the manual system, in terms of the three performance measures used.

INTRODUCTION

The planning of train schedules, which is a crucial task in rail operations planning, is an optimization problem and its complexity has been addressed in various publications. For a single-line network, one solution methodology is to resolve the set of existing conflicts on allocating track sections to the competing trains. It can be shown that the size of solution space grows exponentially as the number of conflicts increase [1]. To resolve a conflict which involves two trains, either of the two may get precedence. Accordingly, for a problem including \( n \) conflicts, the number of solution possibilities is \( 2^n \).

Traditionally, train scheduling experts manually construct the timetables using a distance-time graph, which shows the progress of any train toward its destination with respect to time. It also represents the potential train conflicts and possible decisions. To facilitate the scheduling process and to increase financial returns from huge investments made in the rail industry, various classes of the problem have been investigated and a number of practical decision support systems have been developed and used in real environments. The goal of such systems would be to create appropriate timetables, investigate the robustness of schedules in the presence of unforeseen events and to analyze the capacity of the network with reference to a given transport demand. The simulation approach appears to have caught the attention of researchers and service-planning professionals extensively and has been embraced relatively well by the industry. A few simulation tools have been developed and used in practice. For example, VISION, SIMON, SIMU-PLAN and CAPTURE/PROVING [2-5] have been used for timetabling, TTS and LCAS [6,7] for capacity analysis and Dispatch Analysis Model [8] for both purposes. A general modeling framework has been proposed in [9] for use in simulating rail operations. The complexity of the problem and the possibility of simulating detailed aspects of the problem have contributed to the invention and use of train dispatching rules.

The train scheduling problems have also been formulated using analytical or OR-based techniques, including (non-) linear (mixed) integer programming [10-12] and network optimization models [13,14]. Amongst the solution techniques developed to solve the problem are branch & bound [1], heuristics [15], Lagrangian relaxation [12] and the shortest path method [11, 13]. OR-based solution techniques have provided valuable insights into the problems. However, few models have solved ‘real-life’ problems [9] and succeeded in practice. The reasons could be in the simplified assumptions, the long computing time required and the difficulties in including the domain’s knowledge. Promising
works in the field of artificial intelligence encourage further research in the field to solve the problem more efficiently. Examples are the works of [16,17] in the field of knowledge-based systems and [18] in constraint satisfaction methods. Additionally, AI-based search techniques have been developed by [19-21] and the combinations of knowledge-based techniques with simulation and OR-based methods are presented in [22,23].

This paper presents a hybrid train scheduling algorithm, HeuSA, which is a combination of a simulated annealing control strategy and the constraint-based heuristic introduced in [20]. The hybrid model is tested using the data of manually created timetables and the rail network of Iran. The goal is to find appropriate solutions in a short span of time through starting the search process from a good solution and reducing the likelihood of being trapped in local optima. In addition, new performance measures can easily be used with the model.

In the remainder of the paper, first, the problem at hand is represented and brief discussions on simulated annealing and constraint propagation is provided. Then, insights into the hybrid algorithm developed are provided and simulation results are presented. Finally, concluding remarks are provided.

PROBLEM REPRESENTATION

The problem consists of generating a predictive schedule for a set of passenger trains traveling on a primarily single-line railway that includes some double-line sections, using Iran’s rail network and train data. For any train being considered, the following data is specified by experts: Ready time for boarding at the origin station, minimum stop time considered for reasons such as boarding and alighting at the origin and any intermediate station and point-to-point running time on the track between two adjacent stations, called a block. Data of train operations in the past can be used to work out the above train attributes. Decision variables associated with a train are departure times at the origin and en route stations and arrival times at the destination and en route stations. Thus, stop time at each station is taken as a dependent variable with a value equal to, or greater than, a minimum time planned. The problem constraints are, as follows:

1. For each train, $t$, traveling between two adjacent stations, namely from $s$ to $s+1$:

   $\text{Arrival}_{t,s+1} - \text{Departure}_{t,s} = \text{Running time}_{s,s+1}.$

   (1)

2. For each train, $t$, at each station, $s$, en route:

   $\text{Departure}_{t,s} - \text{Arrival}_{t,s} \geq \text{Planned stop time}_{t,s}.$

   (2)

   Planned stop time is a constant greater than, or equal to, zero.

3. For any pair of trains, $t_1$ and $t_2$, traveling in the same direction, either on a single or double-line track (Figure 1: a, b, c) from station $s$ to its immediate adjacent station, $s+1$:

   $\text{Departure}_{t_1,s} - \text{Arrival}_{t_2,s+1} \geq 0,$

   or:

   $\text{Departure}_{t_2,s} - \text{Arrival}_{t_1,s+1} \geq 0.$

   (3)

   For any pair of opposite trains, $t_1$ and $t_2$, traveling on a single-line track between immediate adjacent stations, $s$ and $s+1$, if $t_1$ is traveling from $s$ to $s+1$ (and $t_2$ from $s+1$ to $s$):

   $\text{Departure}_{t_1,s} - \text{Arrival}_{t_2,s} \geq \text{Safety interval},$

   or:

   $\text{Departure}_{t_2,s+1} - \text{Arrival}_{t_1,s+1} \geq \text{Safety interval},$

   (4)

   where the safety interval is the time interval between consecutive occupation of a single-line track by opposite trains (in this model, safety interval = 1 minute) (Figure 1: d, e and f).

![Figure 1. A distance-time graph for two trains traveling in the same direction in three situations: (a) A resource conflict exists; (b and c) The conflict is resolved using one of the above constraints; (d, e and f) Similar graphs have been drawn for opposite trains. In this type of graph, consecutive departure and arrival times of each train are connected by a line to represent train movements and stops.](image)
4. For each train, \( t \), at its origin station, \( o \):

\[ \text{Departure}_{t,o} \geq \text{Ready time}_{t,o} + \text{Stop time}_{t,o} \]  

The following safety constraints are added to eliminate the possibility of collisions at stations. These constraints ensure that two opposite trains, \( t_1 \) and \( t_2 \), approaching do not enter station \( s \) simultaneously.

\[ \text{Arrival}_{t_1,s} \geq \text{Arrival}_{t_2,s} + \text{Safety interval} \]

or:

\[ \text{Arrival}_{t_2,s} \geq \text{Arrival}_{t_1,s} + \text{Safety interval} \]

5. Each station can accommodate trains using two kinds of track, i.e., platforms and non-platforms; passenger trains stopping for boarding and alighting are planned to use platforms for the comfort of passengers and trains stopping for other reasons are routed to other lines. However, generating a tight schedule for a set of long-distance trains on a single-line railway may make the violation of such a rule inevitable, let alone the consequences of any unforeseen events in real life. For any station with \( p \) platform and \( n \) non-platform lines at any point in time, \( r \), within the scheduling horizon:

- The no. of trains stopped on platforms \( \leq p \).
- The no. of trains stopped on non-platforms \( \leq n \).

In cases where the above line assignment rule has to be overridden, the total number of trains stopped must be less than, or equal to, \( p + n \).

The definition of a performance measure (mathematical objective function) can help measure the quality of any solution and compare solutions generated by different scheduling methods. Such measures are also used for directing moves across solution space looking for higher quality solutions. In this work, the term "imposed waiting time" is used to refer to a delay imposed on a train at a station, due to the occupation of the requested section of main line by another train. When departure time at the origin and running time on the track are predetermined, the scheduling aim could be to distribute the imposed waiting time amongst involved trains, with respect to an appropriate criterion. In this paper, SWWT, AUWT and MRWJ measures are used.

**SWWT (The Sum of Weighted Waiting Time)**

In the real world, trains are of different value in the eyes of passengers and managers of a railway system. Therefore, each train can be assigned a weight and the objective of a scheduling system may be to reduce the SWWT, as represented below:

\[
\text{Minimize } \sum_{i=1}^{\text{Trains}} \left( \text{Weight}_i \times \sum_{\text{Stations}} \text{Imposed waiting time} \right).
\]

**AUWT (Average Unit Waiting Time)**

The spread of delays, with respect to journey duration, appears to be important where fast, slow, long- and short-distance trains are involved. For a given journey, a fast train is expected to face less delay than a train with lower speed. This is justifiable with regard to the investment made to provide a faster service and the higher fares paid by the customers. Plus, a short-distance train may not be expected to face the same amount of delays as a long-distance train. Therefore, the aim could be to minimize AUWT as represented below, where unit waiting time is the ratio of waiting time to minimum travel time for each train being considered.

\[
\min \left[ \frac{\text{Waiting time}}{\sum_{\text{Tracks}} \text{Running time} + \sum_{\text{Stations}} \text{Planned stop}} \right].
\]

The drawback of AUWT is the possibility of generating very good schedules for most trains at the expense of generating very bad schedules for a few others.

**MRWJ (Maximum Ratio of Waiting Time to Journey Duration)**

MRWJ is better when allocating the same amount of waiting time to a long-distance train, where a short-distance train is undesirable and the problem includes both types of train. MRWJ is represented as:

\[
\min \left[ \max_{\text{Trains}} \frac{\sum_{\text{Stations}} \text{Waiting time}}{\sum_{\text{Tracks}} \text{Running time} + \sum_{\text{Stations}} \text{Planned stop}} \right].
\]

For a given schedule, the lower the maximum ratio of
delay to journey duration, the higher the quality of the schedule would be.

**SOLUTION TECHNIQUES USED**

**Simulated Annealing Algorithms**

This technique refers to a process which simulates the process of cooling a collection of hot vibrating atoms. The basic annealing algorithm was proposed by Metropolis, Rosenbluth, Rosenbluth, Teller & Teller [24], while simulated annealing was initially developed by Kirkpatrick, Gelatt & Vecchi [25] for solving combinatorial optimization problems, which yielded promising results. This technique originates from the conjunction between the annealing process and certain characteristics of the resulting substance. It is believed that, in practical contexts, the state of a substance at low temperature is remarkably affected by the annealing process and, thus, a slow cooling process is recommended.

The algorithm is a search control strategy for finding the best, or nearly the best, solution for a problem without exhaustive navigation of the solutions space. It is intelligent in the sense that it is capable of escaping local optimums and plateau and it is not a brute force search technique. SA (Simulated Annealing) algorithms operate under the assumption that a neighborhood can be constructed for identifying adjacent solutions, which can be reached from a solution already generated. The neighborhood mechanism depends on the nature of the problem being treated and the way it is modeled. For example, pair-wise exchanges (or swaps) are frequently used in connection with permutation problems to create a neighbor solution [26]. In a basic algorithm, the search process gets started from an initial solution generated randomly or by a heuristic and continues in a loop through exploring a randomly selected solution in the neighborhood of the current solution at any iteration. For a minimization problem, the move to a better solution (downhill move) is made automatically and the move to an inferior one (uphill move) probabilistically, with the probability of $e^{-\Delta/T}$, where $T$ is called temperature and $\Delta =$ (new solution’s objective function value-current solution’s value). This is a simple, general probabilistic guidance strategy for avoiding the local search descent disadvantage [27]. $T$ is initially set to an appropriately high value to facilitate the selection of low quality solutions in early stages. Then $T$ declines gradually, by cooling rate $\alpha$, where $0 < \alpha < 1$, to reduce the probability of accepting lower quality solutions and converge into better ones. Also, the number of iterations using the same $T$ value increases gradually by $\beta \geq 1$ to allow more effort with lower $T$ values, hoping to achieve further improvement in later stages of the search process. The process stops when $T$ reaches a value close to zero.

**Constraint Satisfaction Problems**

A Constraint Satisfaction Problem (CSP) involves a finite set of variables, where each variable has a finite domain and a set of constraints restricts the assignment of values to the variables. The primary aim of solving the problem is to find a solution, which satisfies the constraints involved. In this context, constraints can contribute to the efficiency of a solution method in different ways. A major contribution of the constraints is to problem modeling; constraints make sure that the encoding of a correct model of the problem is indeed a step towards an efficient running programme. Another contribution of the constraints is propagation, which spontaneously produces the consequences of a decision [28]. Constraint propagation may be viewed as a way of using constraints to reduce the computational efforts required to solve a problem, where constraints are not used only for testing the validity of solutions, but, instead, are used to detect inconsistencies and reduce the domains of variables. For example, if variables $x$ and $y$ denote integers, from $(x < y)$ and $(x > 10)$, one can deduce that the value of $y$ is at least 12 and smaller values must be removed from its domain. If constraint $(y \leq 11)$ is added later, a contradiction is detected. Without propagation, the test $(y \leq 11)$ cannot be performed before the instantiation of $y$.

There are certain methods which use constraint propagation to increase the efficiency of search techniques for solving hard combinatorial problems; for example, look-ahead techniques propagate constraints after labeling any variable by assigning a value, in order to reduce the domains of unlabelled variables and to detect if the problem is insoluble [29]. Once a variable is labeled, domains of unlabelled variables are pruned by removing values, which are inconsistent with the values already allocated to labeled variables.

**"HeuSA", THE HYBRID MODEL**

The scheduling model, HeuSA, represents a combination of simulated annealing with a constraint-based heuristic (CBH) and constraint propagation technique. First, a feasible solution is generated using CBH, then, the neighborhood exploration process starts, according to SA strategies. In fact, the CBH creates the first feasible solution from where the search process gets started. To create a solution, it resolves conflicts between trains requesting a block, according to the earliest time they leave the block; thus, the train which releases the track earlier is given priority to occupy it. Ties are broken, according to the earliest time they are ready to occupy the block, i.e., the train with minimum
start time gets precedence. In this process, the station
capacity constraints are also taken into account. CBH
tends to minimize total local waiting time over all
trains involved.

What is the neighborhood concept of SA in this
context? For a given feasible schedule, a neighbor
solution is a schedule in which the sequence of occup-
ing a block by two requesting trains with overlapping
time intervals is changed, and the destructed part of
the schedule resulting from the swap is identified and
rescheduled using the same heuristic.

A move in the neighborhood of a solution is
conducted using the following steps:

1. Identifying a potential move randomly, which is a
   train delayed at a station for another moving train
to finish the occupation of the requested block;

2. Swapping the sequence of activities (occupying the
   block by the two trains competing); the earliest
   start time of the two activities are updated and
   constraints are posted for propagation;

3. Propagating constraints to identify the activities
   whose start times must be updated; this process
could reveal violations of the resource constraints
   for the part of the schedule affected by the swap-
   ping process. (A look-ahead propagation method is
   used. Further discussions are provided in [20]);

4. Repairing the schedule to modify the earliest start
times of the activities affected by the swap;

5. Evaluating the quality of the new schedule in terms
   of the performance measure selected;

6. Determining the new current schedule using the
   simulated annealing move strategy. Checking the
   threshold to stop the algorithm appropriately. The
   search process continues as long as $T$ is greater
   than a limit close to 0. However, the algorithm stops
   if the number of iterations exceeds a predefined large
   number (i.e., 5000 used in this work).

SIMULATION EXPERIMENTS

The passenger trains scheduling model has been tested
using the data of two corridors of Iran’s railways. Data
are working timetables of passenger trains generated by
train planning experts for Tehran-Mashad and Tehran-
Tabriz, connecting the capital to two major cities in the
northeast and northwest of the country, respectively.
The two routes are transit lines connecting Turkey in
the northwest to Turkmenistan in the Northeast. Five
cases have been considered in this work, where a case
includes a predictive schedule produced manually for
a fleet of trains. The schedule specifies origin and
destination stations, departure time from the origin,
running time between immediate adjacent stations en
route and stop time at any station en route for rea-
sons such as boarding, alighting and meeting opposite
trains. In practice, the schedule covers about 24 hours
and it is repeated daily or weekly. Table 1 shows further
details of the cases solved.

Tables 2 to 4 compare schedules generated by
HeuSA and CBH with manual timetables in terms
of SWWT, AUWT and MRWJ. Figures associated
with SWWT are in terms of minutes and the fig-
ures associated with AUWT and MRWJ have been
multiplied by 100 to make them more understandable.
Table 3. Comparison on the AUWT.

<table>
<thead>
<tr>
<th>Case</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual timetable (MT)</td>
<td>3.27</td>
<td>2.23</td>
<td>2.86</td>
<td>1.97</td>
<td>2.40</td>
</tr>
<tr>
<td>Constraint-based heuristic (CBH)</td>
<td>2.42</td>
<td>1.98</td>
<td>2.22</td>
<td>1.43</td>
<td>2.05</td>
</tr>
<tr>
<td>HeuSA</td>
<td>2.31</td>
<td>2.02</td>
<td>2.16</td>
<td>1.43</td>
<td>2.06</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBH over MT</td>
<td>25.99</td>
<td>11.21</td>
<td>22.38</td>
<td>27.41</td>
<td>14.58</td>
</tr>
<tr>
<td>HeuSA over CBH</td>
<td>4.55</td>
<td>-2.02</td>
<td>2.70</td>
<td>0.00</td>
<td>-0.49</td>
</tr>
<tr>
<td>HeuSA over MT</td>
<td>29.36</td>
<td>9.42</td>
<td>24.48</td>
<td>27.41</td>
<td>14.17</td>
</tr>
</tbody>
</table>

Table 4. Comparison on the MRWJ.

<table>
<thead>
<tr>
<th>Case</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>P4</th>
<th>P5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual timetable (MT)</td>
<td>7.13</td>
<td>4.74</td>
<td>5.38</td>
<td>4.70</td>
<td>5.80</td>
</tr>
<tr>
<td>Constraint-based heuristic (CBH)</td>
<td>6.58</td>
<td>4.54</td>
<td>5.26</td>
<td>3.67</td>
<td>5.69</td>
</tr>
<tr>
<td>HeuSA</td>
<td>5.13</td>
<td>4.51</td>
<td>4.65</td>
<td>3.60</td>
<td>4.87</td>
</tr>
<tr>
<td>Improvement (%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CBH over MT</td>
<td>7.71</td>
<td>4.22</td>
<td>2.23</td>
<td>21.91</td>
<td>1.90</td>
</tr>
<tr>
<td>HeuSA over CBH</td>
<td>22.04</td>
<td>6.66</td>
<td>11.60</td>
<td>1.91</td>
<td>14.41</td>
</tr>
<tr>
<td>HeuSA over MT</td>
<td>28.05</td>
<td>4.85</td>
<td>13.57</td>
<td>23.40</td>
<td>16.03</td>
</tr>
</tbody>
</table>

in percent. Since random variables are incorporated into the neighborhood exploration, each case is solved using six random generator seeds and the average value is reported for each objective function. Simulation experiments show that all solutions created by CBH and HeuSA are better than the manual ones; however, the degree of improvement varies with case and objective function. Test results for CBH show that SWWT has decreased by between 5.04% and 27.34%, AUWT by between 11.21% and 27.41% and MRWJ by between 1.90% and 21.91%. On average, improvements achieved by HeuSA are between 11.09% and 27.34% for SWWT, 9.42% and 29.36% for AUWT and 4.85% and 28.05% for MRWJ.

In three out of thirty experiments (five cases, three objective functions, two algorithms, i.e. CBH and HeuSA), HeuSA did not make any improvement over CBH, although solutions were, in terms of objective function value, only slightly different. In fact, the probabilistic search process started from a very good solution generated and navigated solution space by CBH and no improvement achieved by SA. SA can be tuned to progress the search process more carefully and for a longer period of time to outperform the CBH.

Table 5 shows search efforts made by HeuSA in terms of CPU time and the number of visited solutions, which is represented by the number of iterations. CBH created the initial solution for each case in less than two seconds. Experimental results reveal that between 4.85% and 29.36% improvements have been achieved, whilst consuming 48 to 228 seconds CPU time of a personal computer of type COMPAQ Pentium. The figures in the above tables show that the HeuSA outperforms the manual system in terms of the performance measure value and schedule production time. HeuSA also outperforms CBH, since it yields better solutions in 12 out of 15 tests and the gap is very narrow in the three other tests. From a managerial viewpoint, improvements on SWWT, AUWT and MRWJ appear well enough to encourage the use of the scheduling methods presented as the advisory system for the scheduling of passenger trains.

As explained previously, key algorithm knobs are T, α and β. The initial T value in the model is equal to (−f(Q)/log(0.6)), where f(Q) denotes the performance measure associated with the first solution from where the search started. In fact, the probability of making an uphill move to a solution with f(Q') = 2 × f(Q) is set to be 0.60. The above formula is used to limit the search effort, taking advantage of the high quality of the first solution created by the heuristic. Another advantage is that the initial T value is automatically calculated for any problem being solved and the performance measure used.

Table 5. Performance in terms of CPU time and iterations.

<table>
<thead>
<tr>
<th>Case</th>
<th>SWWT</th>
<th>AUWT</th>
<th>MRWJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CPU</td>
<td>CPU</td>
<td>CPU</td>
</tr>
<tr>
<td>P1</td>
<td>1005</td>
<td>228</td>
<td>543</td>
</tr>
<tr>
<td>P2</td>
<td>864</td>
<td>110</td>
<td>307</td>
</tr>
<tr>
<td>P3</td>
<td>924</td>
<td>181</td>
<td>531</td>
</tr>
<tr>
<td>P4</td>
<td>741</td>
<td>72</td>
<td>462</td>
</tr>
<tr>
<td>P5</td>
<td>879</td>
<td>128</td>
<td>519</td>
</tr>
</tbody>
</table>
Table 6 shows the values allocated to $\alpha$ and $\beta$ for SWWT, AUWT and MRWJ. Typically, $\alpha$ is chosen in a range between 0.90 and 0.99 [30]. The initial values of $\alpha$ and $\beta$ used to start the experiments for tuning the HeuSA, were 0.95 and 1.05 respectively. The final values are shown in Table 6. The table also shows the number of iterations for each $T$ during the search process.

**CONCLUSIONS**

This paper represented a combination of the simulated annealing and a constraint-based heuristic. Starting the search process from a good solution created by the heuristic, the algorithm improved solution quality while avoiding exhaustive search and reducing the possibility of being trapped in local optima. Train scheduling experts were able to run CBH or HeuSA, using the objective function desired. HeuSA was fast enough to encourage users to modify input data to enhance quality. This is the practical advantage of the automated advisory system introduced; this encourages further research to devise interactive intelligent techniques capable of tackling real world problems. Data mining can also be used to extract the knowledge of scheduling experts to create an initial solution of higher quality. In fact, technical experts who create timetables manually may find it very difficult to explain the heuristics they use or might be reluctant to transfer their expertise; in such cases, existing train schedule databases, created in the past, can be explored using data mining techniques to learn about rules and facts used by experts.

**REFERENCES**


