MULTI-STRATEGY REINFORCEMENT LEARNING SIMULATION BASED ON VALUE MODEL FOR PETROLEUM SUPPLY CHAIN

XINYI YANG [1], LIANG SUN [1], BOCHEN YIN [1],
QIAOZHEN QIN [1], ZHE WANG [1], NAN MA [1]
[1]Industrial Chain Optimization Research Center, PetroChina Planning and Engineering Institute, Beijing, China
E-MAIL: yangxy234@petrochina.com.cn, sunliang_cppei@petrochina.com.cn, yinbochen@petrochina.com.cn,
qinqiaozhen@petrochina.com.cn, wang_zhe@petrochina.com.cn, manan2013@petrochina.com.cn

Abstract:

The petroleum supply chain is a sequential process involving various steps. Operational research (OR) is the traditional approach to resolve resource scheduling problem in such a system with complicated structure. However it is time consuming due to enormous amount of variables. In this paper we have managed to build a multi-agent reinforcement learning (RL) environment to simulate decision-making processes, which employs a multi-strategy approach to optimize agent interactions while leveraging value models to enable agents to adapt and improve their actions. Through a series of simulated scenarios, we demonstrate the effectiveness of the proposed multi-strategy framework in addressing complex tasks within multi-agent systems. The framework is designed to maximize overall performance by employing tailored RL strategies.

Keywords:

Petroleum supply chain; Reinforcement learning; Simulation; Multi-strategy framework; Multi-agent; Value model

1. Introduction

The petroleum supply chain is a complicated sequential process that involves various steps that are often separated into three main segments: upstream where crude oil exploration and production takes place, midstream where crude oil is refined in refineries and distributed to the storage of refined oil, and downstream where refined oil is sold [?]. Nowadays the resource scheduling problem in the petroleum area has drowned in massive attention due to complicated structure and considerable scale of the system. Yet traditional approaches such as OR are computationally inefficient, especially for such a large model with an enormous amount of variables.

To resolve the problem of OR methods, we turn to constructing a simulation environment for the RL model based on a multi-strategy framework while introducing the concept of value model which reflects the value of inventory of each storage with respect to the level of stocks.

As a result, we observe that the output of simulation model meets the expectation of inventory trends, which is logically reasonable according to our business experience.

2 Problem Definition and Algorithm

We aim to derive the simulation environment based on multi-agent and multi-period sequential deduction framework where the decision-making process of ”observation-decision-execution-observation” is introduced. With the inventory value model which makes the inventory level trends conform to the business rules strategically oriented, the allocation system with continuous improvement ability is realized.

2.1 Task Definition

Due to profound stability of upstream business operation process, we mainly focus on the downstream which starts from refined oil being produced in refinery, then transported to transit depots, and finally retailed in sales company. Given inventory level of each node (namely refineries, transit depots, and sales companies) and monthly production plan of refineries and sales plan of sales companies as input, we seek to simulate the downstream scheduling process under the condition that the inventory level of each node maintains in certain range steadily as our desire while executing the operational plans smoothly. According to the topology network, each transportation is uniquely defined by four elements: start node, end node, transporting material, and shipping method, the oil can only be delivered along the sequential stream.

2.2 Algorithm Definition

In the RL model, we construct the nodes as agents that interact with the environment and make decisions based on a multi-strategy framework, where each strategy determines shipment to the associated downstream depots. The state space is defined as the union of static transportation network, supply and demand, inventory level, as well as a virtual value varying with the inventory level for each node. The discrete action space is yield by different selection of strategies, value models, and parameters that affect the actions.

2.2.1 Variable Definition

Suppose there is a node j at state s_i, with $v_j^{(k)}$ the k-th value function representing the inventory level of node j and p_k the adjustable parameters of the strategy. The shipment quantity vector from node j to all N downstream nodes is $a^j = [a_1^j, a_2^j, \dots, a_N^j]$. Then the full action space $A_i^{(j)}$ for node j in state s_i is defined as formula 1.

$$A_i^{(j)} = \Big\{ a^j \mid a^j = \pi_k(s_i, v_j^{(l)}, p_k), \quad k \in \{1, \dots, K\},\ l \in \{1, \dots, L\} \Big\} \tag{1}$$

where K and L are the total number of strategies and value functions respectively. In this paper, $K = 5$ and $L = 4$.

2.2.2 Reward Function

The reward function is developed to maximize the change of total value from previous state to the next one. The process is listed as follows.

- Consider a node j with associated value v_j. The total environment value at state s_i is given by formula 2.

$$V(s_i) = \sum_{j=1}^{n} v_j^{(i)}. \tag{2}$$

- When an action a is taken, the environment transitions from state s_i to state s_{i+1}, and each node updates its value, leading to a new total value (formula 3)

$$V(s_{i+1}) = \sum_{j=1}^{n} v_j^{(i+1)}. \tag{3}$$

- The reward function aims to maximize the change in the total value as formula 4

$$R(s_i, a) = V(s_{i+1}) - V(s_i). \tag{4}$$

- The agent is incentivized to choose actions that maximize $R(s_i, a)$, i.e., actions leading to the greatest positive change in the environment value.

The Figure 1 illustrates how the simulation environment takes actions:

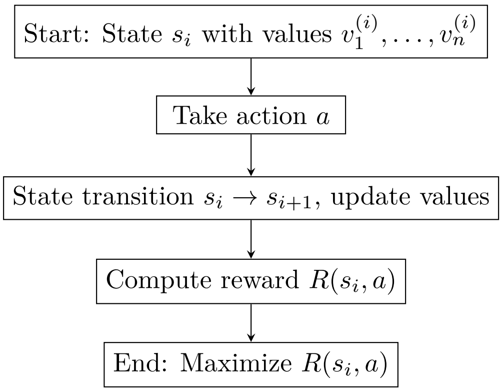
Start: State s_i with values $v_1^{(i)}, \dots, v_n^{(i)}$
Take action a
State transition $s_i \to s_{i+1}$, update values
Compute reward $R(s_i, a)$
End: Maximize $R(s_i, a)$

FIGURE 1. State Transaction and Reward Computation

2.2.3 Multi-Strategy Framework

We use a multi-strategy framework where each strategy determine generates different shipment scenarios based on node status. The 5 strategies we use are listed below.

- Balanced strategy: evenly distributes the transportation volume to all eligible downstream nodes, i.e. the shipment quantity from node i to all N downstream node j is $a_1^j = a_2^j = \cdots a_N^j$ for all $j \in \{1, \dots, N\}$.
- Greedy strategy: gives priority to the downstream node with the largest value difference for transportation to maximize the current value difference each time, so as to achieve the purpose of global optimization in gradual accumulation. This strategy leads to find the downstream node j that maximize $|v_i - v_j|$ for $j \in \{1, \dots, N\}$.

- Balanced-weight strategy: calculate the shipments based on allocation ratio and the actual remaining capacity of the depot using Softmax function. The calculation process for node i with N downstream nodes is as follows:
  1. Calculate current inventory ratio as: $t_i = \frac{c_i}{m_i}$, where c_i and m_i represents current inventory and capacity respectively.
  2. List all downstream nodes that satisfy $v_j >= v_i$, and calculate value difference $\Delta v_i^j = v_i - v_j$ for $j \in \{1, \ldots, N\}$.
  3. Calculate remaining capacity based on current inventory level and total capacity of each downstream nodes as: $r_j = m_j - c_j$ for $j \in \{1, \ldots, N\}$.
  4. Calculate total shipment quantity based on inventory level:
  $$t_i^j = v_i \times \left(\frac{\sum_{j=1}^{N} r_j}{\sum_{j=1}^{N} m_j} \right) \times \min(c_i, \sum_{j=1}^{N} r_j) \tag{5}$$
  5. Calculate standardized value difference u_i^j:
  $$u_i^j = \frac{\Delta v_i^j}{\sum_{j=1}^{N} \Delta v_i^j} \tag{6}$$
  6. Calculate allocated shipment quantity by Softmax:
  $$p_i^j = \frac{e^{u_i^j}}{\sum_{j=1}^{N} e^{u_i^j}} \tag{7}$$
  7. The shipment quantity T_i^j from node i to associate downstream node j is
  $$T_i^j = \min(t_i^j p_i^j, r_j) \tag{8}$$
- Proportional allocation strategy: adopts the method of direct allocation according to the inventory proportion in the distribution process instead of Softmax function compared to balanced-weight strategy. The total shipment quantity is computed as
$$T_i^j = \frac{r_i t_i^j}{\sum_{j=1}^{N} r_j} \tag{9}$$
which simplifies the allocation process while ensuring reasonable transportation.
- Target strategy: adjusts the transportation volume based on the target inventory ratio of each node.This strategy allows us to fully control the inventory level as the result of shipment activities. Same as before, we assume shipment from node i to the j^{th} downstream node and the calculation process is as follows:
  1. Let c_i, g_i, h_i, e_i be current inventory, production, sales, and expected inventory ratio of node i respectively. Then we have equation 10 and 11.
  $$d_i = c_i + g_i - e_i m_i, \tag{10}$$
  $$q_j = (m_j - c_j) + h_j - (1 - e_j) m_j \tag{11}$$
  where d_i is the delivery quantity of node i and q_j is the receiving quantity of node j.
  2. The total shipment quantity of node i is
  $$T_i = \min \left(\sum_{j:v_i<v_j,q_j>0} q_j, d_i \right) \tag{12}$$
  3. Then we can allocate each shipment quantity from node i to node j by
  $$T_i^j = \frac{T_i q_j}{\sum_{j:v_i<v_j,q_j>0} q_j} \tag{13}$$

3 Experiment

We mainly discuss the outcome of the experiments using balanced-weight strategy and linear value model with different original inventory and parameter settings.

3.1 Settings

Take the data of an arbitrary month as an example, we construct the RL model including 90 nodes, 2 types of oil and more than 1000 routes in the transport network. Monthly production and sales plans sum up to 50.

We use linear value model in the experiment as shown in formula 14, with v as the inventory rate, v_0 and v_{100} as parameters representing values at empty and full inventory respectively.

$$\text{Value} = (100 - v) \times \frac{(v_0 - v_{100})}{100} + v_{100} \tag{14}$$

In the experiment we set $v_0 = 0$ for all nodes, and $v_{100} = 1000$ for refineries and transit depots, and $v_{100} = 1500$ for sales companies.

3.2 Results

The following examples are using the same dataset except subtle adjustments on a specific refinery (denoted R) of interest.

3.2.1 Base Example

According to real operational data, the inventory ratio at the beginning of the month is 65% and 50% for gasoline and diesel respectively. With shipping threshold set to 2000 tons. Figure 2 shows the trend of inventory level and value of refinery R through the whole month as the simulation outcome. Due to shipping threshold, shipments begin on the third day, then inventory level and value oscillate in the opposite direction.

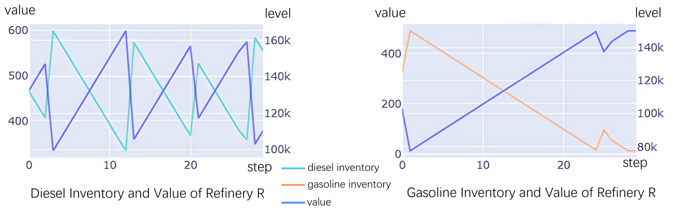

FIGURE 2. Trend of inventory level and value of refinery R (Basic Example)

3.2.2 High-Initial-Inventory Example

Let the initial inventory of both gasoline and diesel be 100% with shipping threshold 1000 tons (Figure 3). It can be seen that oil is delivered out of refinery R in a very short time, decreasing the inventory to a reasonable level and stabilized. The shipments are triggered relatively more often compared to base example.

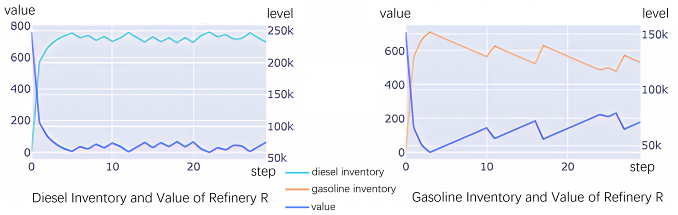

FIGURE 3. Trend of inventory level and value of refinery R (High-Initial-Inventory Example)

3.2.3 Low-Initial-Inventory Example

Now change the initial inventory of both gasoline and diesel to empty and keep shipping threshold unchanged (Figure 4). It is obvious that transportation activities are allowed when the inventory is low, and only begin until sufficient quantities have been produced and accumulated.

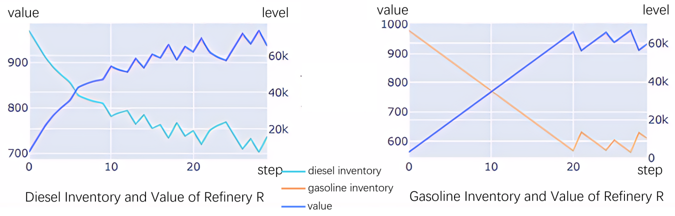

FIGURE 4. Trend of inventory level and value of refinery R (High-Initial-Inventory Example)

3.3 Discussion

Based on three examples illustrated in detail above, we conclude that:

- The linear value model algorithm based on the current business logic tends to stabilize the inventory within a reasonable range. The stocks are quickly transferred at high inventory level, and accumulated when the inventory is low, so as to automatically adjust the inventory level.
- The shipping threshold can affect the fluctuation range of the inventory trend in a way that lower threshold brings less inventory fluctuations hence flatter pattern and vice versa.
- The BWSA strategy aims to narrow the gap between all possible transportation volumes preventing the excessive proportion gap from causing some nodes unable to receive oil due to the too small transportation volume.

4 Related Work

The structure of petroleum supply chain is particularly critical due to the scale and global reach of operations. Managing a complex and distributed crude oil supply chain poses significant computational challenges, particularly when using traditional OR algorithms. These methods, while powerful, often require extensive computational resources and can suffer from long processing times, especially when dealing with large-scale, multi-echelon systems. The complexity of such algorithms increases exponentially with the number of nodes and the stochastic nature of demand and lead times, often leading to scenarios where no feasible solution can be found within a reasonable time frame. This makes them less practical for real-time decision-making in dynamic supply chain environments like those found in the crude oil industry.

To address these significant challenges, researchers in the field of OR have explored various innovative methods. Edirisinghe and Almutairi [?] introduced a predictive global sensitivity analysis (PGSA) approach to simplify computational processes by creating structural equations based on regression techniques while still providing near-optimal solutions. Similarly, Sitek and Wikarek [?] proposed a hybrid framework that combines mathematical programming and constraint programming to optimize decision problems in sustainable supply chain management.

However, existing research has largely focused on simpler models. Previous studies have typically focused on more straightforward two-echelon systems, where a central warehouse supplies multiple retailers [?]. Hearnshaw and Wilson [?] proposed Supply Chain Network Theory that acknowledgs the significance of scale-free networks. The exploration of the simulation model of pharmaceutical supply chain in Morocco only explored the non-distributed inventory system [?].

Despite the various advanced OR methods explored to address the computational challenges in complex supply chains, these approaches often remain limited by their inherent complexity, extensive computation times, and the exponential increase in difficulty as system scale and stochastic factors grow. To address these challenges, more simplified, yet effective, strategies such as the (R, Q) and (S,s) inventory models have been adopted. These models focus on optimizing inventory levels at individual nodes within the supply chain, rather than attempting to optimize the entire network simultaneously. The (R, Q) model, in particular, has been widely used in various industries for its simplicity and effectiveness in managing inventory with uncertain demand, as demonstrated in the context of two-echelon inventory systems [?].

5 Future Work

Current strategy system with value model is constructed solely based on inventory, whereas other potential factors should also be taken into serious consideration, such as transportation duration, delivery cost, storage price, and sales price, etc. The absence of these factors may account for the current system's inability to achieve the lowest possible cost.

Future research should investigate a wider range of strategic frameworks and value models, along with diverse configurations of parameters such as shipping thresholds. The current classification method considers only the positional attribute of nodes in the supply chain, whether they are upstream or downstream, without taking into account their specific functional roles or business impact within the network. A more refined node categorization could be established to support the development of targeted allocation strategies, value models, and parameter settings thereby enhancing the overall effectiveness and precision of the system.

6 Conclusion

This paper presents a RL simulation environment in the area of petroleum supply chain based on multi-strategy framework introducing value model to determine the best approach that meet business rules. In consequence, the model conforms to the business logic by analyzing three examples. Further research on strategies and value models is worth studying given current research progress.