MARKETING METHOD AND APPARATUS BASED ON DEEP REINFORCEMENT LEARNING

TECHNICAL FIELD

[0001] One or more embodiments of the present specification relate to the field of machine learning, and in particular, to marketing methods and apparatuses based on deep reinforcement learning.

BACKGROUND

[0002] With rapid development of machine learning, various machine learning models have been applied to a variety of service scenarios, such as evaluating credit risks of users, recommending appropriate content to users, sending appropriate marketing information to users, or predicting advertisement delivery effects.

[0003] In fact, a complete service process in a service scenario often includes multiple phases or multiple stages, and involves many factors. For example, in a recommendation marketing service scenario, a service can include at least the following stages: selection and sending of marketing information, and user feedback and tracing after the sending. These stages involve many factors and parameters, such as a marketing channel, a sending method, a reach rate, and a conversion rate. Therefore, in attempts to apply machine learning to service scenarios, how to design and select applicable models and what factors need to be introduced to measure service execution results are problems to be considered in the case of many factors and stages in complex service scenarios.

[0004] Therefore, improved solutions are expected to improve service execution effects more efficiently by using machine learning.

SUMMARY

[0005] One or more embodiments of the present specification describe marketing methods and apparatuses based on deep reinforcement learning, so as to simultaneously learn multiple targets in a recommendation marketing service, thereby comprehensively promoting a service effect.

[0006] According to a first aspect, a marketing method based on a deep reinforcement learning system is provided, where the deep reinforcement learning system includes an agent and an execution environment, the agent is configured to determine a marketing behavior based on a marketing strategy and status information of the execution environment, and the method includes the following: obtaining, from the execution environment, multiple execution results generated when a user responds to the marketing behavior, where the multiple execution results respectively correspond to multiple target effects on a marketing effect chain; determining a reward score of reinforcement learning based on the multiple execution results; and returning the reward score to the agent for the agent to update its marketing strategy.

[0007] According to some embodiments, the status information of the execution environment can include alternative marketing information and user information of the current user.

[0008] In some embodiments, the multiple execution results of the user are obtained in the following method: obtaining an interaction behavior of the user from a server, and obtaining the multiple execution results by using the interaction behavior; or capturing an operation behavior of the user by using a client device user interface tracing point, and obtaining the multiple execution results by using the operation behavior of the user.

[0009] In some embodiments, the reward score is determined by using a score evaluation function, where the score evaluation function uses the multiple execution results as variables and is designed to be positively correlated with a match rate between the multiple execution results and multiple corresponding target effects.

[0010] In some other embodiments, the determining a reward score of reinforcement learning includes: obtaining independent functions constructed separately for at least some of the multiple execution results, and determining the reward score based on at least a weighted sum of the independent functions.

[0011] Further, in some embodiments, the independent functions can be constructed to be positively correlated with a corresponding match rate between an execution result and a target effect.

[0012] In some possible embodiments, the multiple execution results include a first execution result, and the first execution result has a discrete result; and the independent functions include a first function, and the first function outputs a discrete value based on the discrete result of the first execution result.

[0013] Furthermore, according to some implementations, the first function outputs a first discrete value when the first execution result is a first result and outputs a second discrete value when the first execution result is a second result, where the first result has a higher match rate with a first target effect corresponding to the first execution result than the second result, and the first discrete value is greater than the second discrete value.

[0014] According to some implementations, the reward score can further be determined in the following method: determining at least one result combination based on the multiple execution results; obtaining at least one combination function constructed for the at least one result combination; and determining the reward score based on at least a weighted sum of the at least one combination function.

[0015] According to some other implementations, the reward score can further be determined in the following method: determining a target effect or a target effect combination that matches the multiple execution results; obtaining a pre-established mapping relationship between a target effect or a target effect combination and a reward score; and determining, based on the mapping relationship, a reward score corresponding to the matched target effect or target effect combination.

[0016] According to a second aspect, a marketing apparatus based on a deep reinforcement learning system is provided, where the deep reinforcement learning system includes an agent and an execution environment, the agent is configured to determine a marketing behavior based on a marketing strategy and status information of the execution environment, and the apparatus includes the following: an acquisition unit, configured to obtain, from the execution environment, multiple execution results generated when a user responds to the marketing behavior, where the multiple execution results respectively correspond to multiple target effects on a marketing effect chain; a determining unit, configured to determine a reward score of reinforcement learning based on the multiple execution results; and a returning unit, configured to return the reward score to the agent for the agent to update its marketing strategy.

[0017] According to a third aspect, a computing device is provided, and includes a memory and a processor, where the memory stores executable code, and when executing the executable code, the processor implements the method according to the first aspect.

[0018] According to the methods and the apparatuses provided in some

embodiments of the present specification, the deep reinforcement learning system is used to learn recommendation marketing services. When updating its marketing strategy, the agent considers multiple target effects on the marketing effect chain, that is, the deep reinforcement learning system simultaneously learns multiple targets on the marketing chain, so as to comprehensively learn an entire service process and service targets of recommendation marketing, thereby better promoting a service execution effect.

BRIEF DESCRIPTION OF DRAWINGS

[0019] To describe the technical solutions in some embodiments of the present specification more clearly, the following briefly describes the accompanying drawings needed for describing the embodiments. Clearly, the accompanying drawings in the following description show merely some embodiments of the present specification, and a person of ordinary skill in the art can still derive other drawings from these accompanying drawings without creative efforts.

[0020] FIG. 1 is a schematic diagram illustrating recommendation marketing execution, according to some embodiments of the present specification;

[0021] FIG. 2A illustrates a typical deep reinforcement learning system;

[0022] FIG. 2B illustrates a deep reinforcement learning-based recommendation marketing system, according to some embodiments;

[0023] FIG. 3 illustrates a marketing method based on a deep reinforcement learning system, according to some embodiments; and

[0024] FIG. 4 is a schematic block diagram illustrating an apparatus, according to some embodiments.

DESCRIPTION OF EMBODIMENTS

[0025] The following describes the solutions provided in the present specification with reference to the accompanying drawings.

[0026] According to one or more embodiments of the present specification, a deep reinforcement learning method is used to promote a recommendation marketing service effect. FIG. 1 is a schematic diagram illustrating recommendation marketing execution, according to some embodiments of the present specification. In

recommendation marketing scenarios, a marketing process can include the following multiple stages. A system first selects, based on some recommendation algorithms, some marketing information suitable for a current user to prepare for recommendation. The marketing information can include an advertisement, a credit package (for example, a red envelope, a shopping voucher, and a coupon), and a marketing notification message, etc. Then, the recommendation marketing information is sent to a marketing delivery platform for delivery. The marketing delivery platform selects, based on predetermined delivery rules and strategies, some marketing information from the recommendation marketing information for exposure. Once a piece of marketing information is exposed, it is considered that the marketing information reaches users. For example, in FIG. 1, when a user opens a client device App (for example, ALIPAY) and sees a certain piece of exposed marketing information (for example, marketing information 110), it is considered that the marketing information reaches the user.

[0027] Then, the user may click the marketing information, for example, click the marketing information 110 to enter a new user interface 120. Afterwards, the user may stay on the new user interface 120 for a period of time, and perform a second click (namely, a secondary jump) on an interested element on the user interface 120. For example, the secondary jump jumps to a new user interface 130. On the user interface 130, the user completes a target operation of marketing content, such as registering, forwarding, and using the credit package in the marketing information for consumption. In this case, it can be considered that user conversion is achieved.

[0028] The above-mentioned description shows a typical user operation sequence, that is, a typical user marketing effect chain, expected in recommendation marketing. The typical user marketing effect chain includes multiple target effects that are expected to be successively achieved, for example, reach –> click –> stay –> secondary jump –> conversion, where user conversion is an ultimate target of recommendation marketing.

[0029] In some implementation solutions, modeling and machine learning are performed for the ultimate target to promote service effects. For example, if user conversion is the ultimate target in recommendation marketing, modeling and learning are performed specific to a user conversion rate. However, ultimate user conversion is usually implemented through multiple times of interaction in multiple steps, and a traffic funnel phenomenon naturally exists in an interaction process, that is, the later

the step, the lighter the traffic. Therefore, if only an ultimate service indicator, that is, the conversion rate, is learned, a relatively small quantity of positive samples can be used for learning.

[0030] In some other implementation solutions, considering that steps in multiple stages of the interaction process have certain impact on the ultimate target, modeling and learning are separately performed on a target effect of each stage. For example, modeling and learning are separately performed for a reach stage and a click stage, etc. to obtain, for example, a reach rate estimation model, a click-through rate estimation model, and a conversion rate estimation model, etc. With such solutions, more samples used for learning can be obtained, and targeted learning can be performed for each stage. However, in these solutions, each model is trained separately, and a cost is high. If an overall service execution effect needs to be estimated, multiple models need to be used for prediction one by one and prediction results need to be combined. Complexity and an operation cost of a prediction process are relatively high.

[0031] Based on comprehensive consideration of the above-mentioned solutions, some embodiments of the present specification further propose a solution that uses a deep reinforcement learning system for recommendation marketing. FIG. 2A illustrates a typical deep reinforcement learning system. Generally, the deep reinforcement learning system includes an agent and an execution environment, and the agent continuously learns and optimizes its strategy through interaction and feedback with the execution environment. Specifically, the agent observes and obtains a state of the execution environment, and determines, based on a certain strategy and the state of the execution environment, a behavior or an action to take. Such a behavior affects the execution environment, changes the state of the execution environment, and generates feedback to the agent. The feedback is also referred to as a reward score. Based on the obtained reward score, the agent determines whether the above-mentioned behavior is correct and whether the strategy needs to be adjusted, so as to update its strategy. The agent can continuously update the strategy by constantly observing a state, determining a behavior, and receiving feedback. An ultimate objective is to learn a strategy, so that the obtained reward score is maximized. The above-mentioned description is a typical reinforcement learning process. During strategy learning and adjustment, if the agent uses some deep learning algorithms including a neural network algorithm, such a system is referred to as a deep reinforcement learning system.

[0032] FIG. 2B illustrates a deep reinforcement learning-based recommendation marketing system, according to some embodiments. To be specific, deep reinforcement learning is applied to a recommendation marketing system. In recommendation marketing scenarios, the execution environment is a recommendation marketing environment, and an observable environment state includes to-be-recommended marketing information and current user information, etc. The agent determines an appropriate marketing behavior based on some marketing strategies and a current state of the marketing environment. The marketing behavior changes the state of the marketing environment and generates a reward score, which is fed back to the agent. The agent can further adjust and update its marketing strategy based on the reward score. According to one or more embodiments of the present specification, a reward score generation process is optimized, and a comprehensive reward score is determined based on multiple execution results (for example, whether a click is performed, whether a secondary jump exists, and stay duration) corresponding to the multiple target effects on the marketing effect chain. As such, an update of the agent's marketing strategy considers the multiple target effects on the marketing effect chain, that is, the deep reinforcement learning system simultaneously learns multiple targets on the marketing chain, so as to comprehensively learn an entire service process and service targets of recommendation marketing, thereby better promoting a service execution effect. The following describes implementations of the above-mentioned idea.

[0033] FIG. 3 illustrates a marketing method based on a deep reinforcement learning system, according to some embodiments. It can be understood that the deep reinforcement learning system can be the deep reinforcement learning-based recommendation marketing system shown in FIG. 2B. The deep reinforcement learning system includes an agent and an execution environment. The agent is configured to determine a marketing behavior based on a marketing strategy and a state of the execution environment, and update the marketing strategy based on a reward score fed back for the marketing behavior. The method can be performed by any apparatus, device, platform, or device cluster, etc. that has data computing and processing capabilities. As shown in FIG. 3, the method includes at least the following: step 31: obtaining, from the execution environment, multiple execution results separately corresponding to a user for multiple target effects on a marketing effect chain, where the multiple execution results are generated when the user responds to

the marketing behavior determined by the agent; step 33: determining a reward score of reinforcement learning based on the multiple execution results; and step 35: returning the reward score to the agent for the agent to update its marketing strategy. The following describes specific implementations of the above-mentioned steps.

[0034] As described above, in the deep reinforcement learning-based recommendation marketing system, the agent continuously observes and obtains a current state of the recommendation marketing environment, and determines, based on the marketing strategy, a marketing behavior to take in the current state. In some implementations, state information of the marketing environment obtained by the agent can include optional marketing information and user information of a current user. More specifically, the optional marketing information can include optional marketing content, such as an alternative advertisement, a credit package (such as a red envelope, a shopping voucher, and a coupon), and a marketing notification message. The optional marketing information can further include optional marketing channels, such as a message sending, an advertisement banner, and a badge reminder. On the other hand, the state information of the marketing environment further includes the user information of the current user, and the user information can include user attribute characteristics, such as a registration time period, an age, a gender, and an income. In some embodiments, the user information further includes the user's historical behavior information, such as a historical operation sequence, a list of recently browsed user interfaces, and a list of recently accepted marketing information. In some embodiments, the user information can include user profile information. It can be understood that, in some implementations, a corresponding user profile is granted to the user based on some algorithms, and basic attribute information and/or historical behavior information of the user. The user profile information can include, for example, a crowd label for classifying the user into a specific crowd based on attribute information, a cluster label for clustering the user, and a user behavior habit label constructed based on historical user behaviors, etc. The various user information can constitute a user feature of the current user.

[0035] In some specific examples, a state vector of the current state of the marketing environment is constructed based on a marketing feature corresponding to the marketing information and the user feature corresponding to the current user, and then is input to the agent. Based on some marketing strategies and the current state of the marketing environment, the agent determines the marketing behavior to take.

[0036] It can be understood that, for the reinforcement learning system, a strategy in the agent is used to map the state of the environment to the next behavior. In the deep reinforcement learning-based recommendation marketing system shown in FIG. 2B, the agent is an entity that makes a decision, and maps, based on the marketing strategy, the current state of the marketing environment to the next marketing behavior to take. The marketing strategy is implemented by using complex deep learning algorithms, such as Q-learning and DQN. These algorithms generally include some deep neural network structures. As described above, the current state of the marketing environment reflects the user feature of the current user and a feature of the optional marketing information. Once obtaining such current state information, the agent can determine, by using a deep learning algorithm, the marketing behavior to take for the current user, for example, determine specific marketing content and a specific channel used to send the marketing content to the user.

[0037] An execution result is generated accordingly after such a marketing behavior affects the execution environment. Therefore, in step 31, the multiple execution results generated for the multiple target effects on the marketing effect chain when the user responds to the above-mentioned marketing behavior can be obtained from the execution environment.

[0038] FIG. 1 shows a marketing effect chain, according to some specific examples. The marketing effect chain includes multiple target effects that are expected to be successively achieved, including reach –> click –> stay –> secondary jump –> conversion. Correspondingly, the execution results of the user for these target effects can include a marketing-to-user reach status (for example, whether the user is reached, a reach time, or a display time), the user's click status (for example, whether a click is performed), the user's stay duration, the user's second-click status (for example, whether a second click, that is, a secondary jump, is performed), and a user conversion status (for example, whether the user is converted).

[0039] In some embodiments, the execution results of the user for the target effects are reflected as interaction between the user and the execution environment, especially interaction between the user and a marketing server. Therefore, the execution results can be obtained by obtaining an interaction behavior of the user from the server. In some other embodiments, an operation behavior of the user can further be captured by using a client device user interface tracing point, and the execution results of the user for the target effects can be obtained by using the

operation behavior of the user.

[0040] For the obtained multiple execution results generated for the multiple target effects when the user responds to the marketing behavior, in step 33, the reward score of reinforcement learning is determined based on the multiple execution results.

[0041] The following describes the determining of the reward score with reference to the corresponding user execution results and the marketing effect chain of reach –> click –> stay –> secondary jump –> conversion.

[0042] In some embodiments, a score evaluation function is constructed by using all of the multiple execution results as variables, and the score evaluation function is designed to be positively correlated with a match rate between the multiple execution results and the corresponding multiple target effects. To be specific, a higher overall match rate between the multiple execution results and the corresponding multiple target effects can indicate a larger output value of the evaluation function, that is, a higher reward score.

[0043] The score evaluation function is as follows: $Rt=f(X1,X2,\ldots,Xn)$ (1)

[0044] X1, X2, ..., and Xn are execution results respectively corresponding to target effects Y1, Y2, ..., and Yn. In addition, the score evaluation function f is designed as follows: A higher match rate between the execution results X1, X2,... , Xn and the target effects Y1, Y2,... , and Yn can indicate a higher Rt score.

[0045] In some other embodiments, corresponding independent functions are separately constructed for at least some of the multiple execution results. During determining of the reward score, these pre-constructed independent functions are obtained, and the reward score is determined based on at least a weighted sum of these independent functions. More specifically, an independent function fi(Xi) is constructed for the execution result Xi of the ith item, and a final reward score is determined as follows: $Rt=Wi*fi(Xi)$ (2)

[0046] Wi is a weight corresponding to the ith execution result.

[0047] For example, in the above-mentioned example of the marketing effect chain, the multiple execution results include X1=display, indicating a reach status; X2=click, indicating a click status; X3=stay, indicating the user's stay duration; X4=click2, indicating a second-click status; and X5=convert, indicating a conversion status. Therefore, the reward score can be determined as follows: $Rt=W1*f1(display)+W2*f2(click)+ W3*f3(stay)+W4*f4(click2)+W5*f5(convert)$ (3)

[0048] In some embodiments, at least one independent function fi(Xi) is

determined to be positively correlated with a match rate between the execution result Xi and the corresponding target effect Yi. To be specific, a higher match rate between Xi and Yi can indicate a larger value of fi(Xi).

[0049] In still some other embodiments, at least one result combination is determined based on the multiple execution results. Each result combination includes two or more execution results. At least one combination function is constructed for the at least one result combination. The reward score is determined based on at least a weighted sum of the at least one combination function.

[0050] For example, in the above-mentioned example of the multiple execution results of X1 to X5, X1 and X2 can be selected as a first result combination, and a first combination function F1 is constructed for X1 and X2; and X3 and X4 are selected as a second result combination, and a second combination function F2 is constructed for X3 and X4. The result combinations can include different quantities of execution results.

[0051] Therefore, in some specific examples, the reward score can be determined as follows: Rt=WF1*F1(display, click)+WF2*F2(stay, click2) (4)

[0052] The present embodiment can be combined with the embodiment corresponding to equation (2) or equation (3). For example, corresponding independent functions are constructed for execution results that are not selected into a result combination, and weighted summation is performed on these independent functions and the combination function. Or, weighted summation on a combination function is further introduced based on weighted summation on functions corresponding to the execution results. Or, some of the multiple execution results are selected to construct corresponding independent functions, and weighted summation is performed on these independent functions and a combination function.

[0053] For example, in some specific examples, the reward score can be determined as follows:

Rt=W3*f3(stay)+W4*f4(click2)+W5*f5(convert)+WF1*F1(display, click)+WF2*F2(stay, click2) (5)

[0054] X3, X4, and X5 are selected from X1 to X5, and independent functions are constructed for X3, X4, and X5. X1 and X2 are selected as a first combination, and a first combination function F1 is constructed. X3, X4 are selected as a second combination, and a second combination function F2 is constructed. Weighted summation is performed on the independent functions and the combination functions

to determine the reward score.

[0055] In some embodiments, the multiple execution results include a first-type execution result, and such an execution result have a discrete result value. For example, X2 can indicate the click status. In some examples, the click status can be divided into binary discrete result values of "clicked" or "not clicked".

[0056] A discrete function can be constructed for such a first-type execution result as its independent function, and the discrete function outputs a discrete function value based on the discrete result value of the execution result. Further, in some embodiments, a discrete function fi is set for a first-type execution result Xi, and a discrete output value of the discrete function fi can be set as follows: A higher match rate between the discrete result value of the execution result Xi and the corresponding target effect Yi can indicate a larger discrete value output by fi. For example, a discrete function f2(X2) can be constructed for X2. When X2 corresponds to "clicked", f2 outputs a discrete value V1, or when X2 corresponds to "not clicked", f2 outputs a discrete value V2. More specifically, because the corresponding target effect is "clicked", X2 corresponding to "clicked" better matches the target effect. Therefore, in some examples, V1 can be set to be greater than V2, for example, V1=1 and V2=0.

[0057] The present embodiment can be combined with any one of the embodiments corresponding to equation (2) to equation (5), to set an independent function specific to the first-type execution result as a discrete function that outputs a discrete value. For example, in combination with equation (3), for X1, X2, X4, and X5 that have binary results in X1 to X5, their corresponding independent functions can be reconstructed as I, which can indicate a function that outputs 0 or 1. In this case, equation (3) can be reconstructed as follows: Rt=W1*I(display)+W2*I(click)+ W3*f3(stay)+W4*I(click2)+W5*I(convert) (6)

[0058] The function I(X) outputs 1 when X is true and 0 when X is false. It can be understood that X being true corresponds to a case consistent with a target effect.

[0059] In some embodiments, the multiple execution results further include a second-type execution result. Such an execution result corresponds to a continuous variable. For example, X3 represents stay duration, and the stay duration can be represented as a continuous variable using seconds as its unit. A second-type function whose output is a continuous value can be constructed for the second-type execution result. In some specific examples, the second-type function can be a logical regression function. In some other specific examples, a linear regression function is constructed

for the second-type execution result.

[0060] For example, for the continuous variable X3, a corresponding function f3 can be constructed as follows: f3(X3)=A+B*X3 (7)

[0061] A and B are predetermined linear parameters.

[0062] It can be understood that the embodiment corresponding to equation (7) can be combined with any one of the embodiments corresponding to equation (2) to equation (6).

[0063] According to some other implementations, a mapping relationship between a target effect or target effect combination to be achieved and a reward score is pre-established. After an execution result corresponding to a target effect is obtained, a target effect combination matched with the current execution result is determined, and further a corresponding reward score is determined based on the mapping relationship.

[0064] In some specific examples, pre-established mapping relationships can be listed in the following table.

Reward score	Target effect/Target effect combination
1	Display
2	Display + click
3	Stay
4	Stay + secondary jump
5	Conversion

Table 1

[0065] As listed in Table 1, for the target effects that are expected to be successively achieved on the marketing effect chain, if only the target "display" or "reach" is achieved, the reward score is 1; if the target effect "click" by a user is further achieved based on "display", the reward score is 2, and so on. A higher match rate between an achieved target effect and the ultimate target on the marketing effect chain can indicate a high reward score.

[0066] Based on such mapping relationships, after an execution result corresponding to each target effect is obtained, comparison and matching are performed between each execution result and the corresponding target effect, so as to determine a target effect or a target effect combination matched with the current execution result, and further, a corresponding reward score can be determined by

searching the mapping relationships in Table 1.

[0067] It should be understood that Table 1 is merely an example. More, less, or different target effects or target effect combinations can be set based on service needs, and different score values can be correspondingly set for the reward scores.

[0068] The above-mentioned description describes multiple methods for determining a reward score with reference to the marketing effect chain of reach –> click –> stay –> secondary jump –> conversion. It can be understood that the marketing effect chain is not limited to the above-mentioned example, but can include more, less, or different target effects. For example, in some marketing scenarios, after a user performs a click, the user either is converted or exits, without requiring a secondary jump. In some other marketing scenarios, a user may further perform a click after a secondary jump, for example, a third jump occurs. In still some other marketing scenarios, a user may perform other interaction, such as entering personal information, after performing a click or performing a secondary jump. These operations can be used as predetermined target effects on a marketing effect chain, and can further be used as objects learned by the reinforcement learning system.

[0069] The reward score of reinforcement learning is determined based on the multiple execution results of the user for the multiple target effects on the marketing effect chain. Therefore, in step 35, the reward score is returned to the agent for the agent to update its marketing strategy.

[0070] As described above, the agent continuously obtains feedback of the reward score from the environment to determine whether the above-mentioned marketing strategy and marketing behavior are appropriate, so as to update the marketing strategy. It should be noted that, updating the marketing strategy here includes not only modifying or adjusting the above-mentioned marketing strategy when it is determined that the delivered marketing behavior is improper, but also enhancing or maintaining an original marketing strategy when it is determined that the marketing behavior has a good effect. In one or more embodiments of the present specification, as the reward score is determined based on the multiple execution results of the user for the multiple target effects, the agent comprehensively considers the execution results for the multiple target effects when updating the marketing strategy, that is, learns the multiple target effects on the marketing effect chain. As such, the deep reinforcement learning system comprehensively learns an entire service process and service targets of recommendation marketing, thereby better promoting a service

execution effect.

[0071] According to another aspect, some embodiments of the present specification further provide an apparatus. The apparatus is applied to a system that performs recommendation marketing by using deep reinforcement learning. As shown in FIG. 2B, a deep reinforcement learning system includes an agent and an execution environment, and the agent is configured to determine a marketing behavior based on a marketing strategy and status information of the execution environment. The apparatus can be implemented by any software, hardware, or hardware and software combination that has computing and processing capabilities. In some embodiments, the apparatus can be integrated into the execution environment of the deep reinforcement learning system. In some other embodiments, the apparatus can be integrated into an entity apparatus for implementing the agent in the deep reinforcement learning system. In still some other embodiments, the apparatus is used as an independent apparatus to extract data from the execution environment for processing to obtain a reward score and return the reward score to the agent.

[0072] FIG. 4 is a schematic block diagram illustrating an apparatus, according to some embodiments. As shown in FIG. 4, the apparatus 400 includes: an acquisition unit 41, configured to obtain, from the execution environment, multiple execution results generated when a user responds to the marketing behavior, where the multiple execution results respectively correspond to multiple target effects on a marketing effect chain; a determining unit 43, configured to determine a reward score of reinforcement learning based on the multiple execution results; and a returning unit 45, configured to return the reward score to the agent for the agent to update its marketing strategy.

[0073] In some embodiments, the status information of the execution environment includes alternative marketing information and user information of the current user.

[0074] In some embodiments, the obtaining unit 41 obtains the multiple execution results of the user in the following method: obtaining an interaction behavior of the user from a server, and obtaining the multiple execution results by using the interaction behavior; or capturing an operation behavior of the user by using a client device user interface tracing point, and obtaining the multiple execution results by using the operation behavior of the user.

[0075] In some embodiments, the determining unit 43 determines the reward score by using a score evaluation function, where the score evaluation function uses

the multiple execution results as variables and is designed to be positively correlated with a match rate between the multiple execution results and multiple corresponding target effects.

[0076] In some other embodiments, the determining unit 43 determines the reward score in the following method: obtaining independent functions constructed separately for at least some of the multiple execution results, and determining the reward score based on at least a weighted sum of the independent functions.

[0077] Further, in some examples, the independent functions can be constructed to be positively correlated with a corresponding match rate between an execution result and a target effect.

[0078] In some examples, the multiple execution results include a first execution result, and the first execution result has a discrete result; and the independent functions include a first function, and the first function outputs a discrete value based on the discrete result of the first execution result.

[0079] In some further examples, the first function outputs a first discrete value when the first execution result is a first result and outputs a second discrete value when the first execution result is a second result, where the first result has a higher match rate with a first target effect corresponding to the first execution result than the second result, and the first discrete value is greater than the second discrete value.

[0080] According to some implementations, the determining unit 43 can further determine the reward score as follows: determining at least one result combination based on the multiple execution results; obtaining at least one combination function constructed for the at least one result combination; and determining the reward score based on at least a weighted sum of the at least one combination function.

[0081] According to some other implementations, the determining unit 43 determines the reward score as follows: determining a target effect or a target effect combination that matches the multiple execution results; obtaining a pre-established mapping relationship between a target effect or a target effect combination and a reward score; and determining, based on the mapping relationship, a reward score corresponding to the matched target effect or target effect combination.

[0082] Because the determining unit 43 determines the reward score based on the user's multiple execution results of the multiple target effects. Therefore, when updating its marketing strategy, the agent comprehensively considers the execution results of the multiple target effects, so that the deep reinforcement learning system

comprehensively learns an entire service process and service targets of recommendation marketing, thereby better promoting a service execution effect.

[0083] According to some embodiments of still another aspect, a computing device is further provided, including a memory and a processor, where the memory stores executable code, and when executing the executable code, the processor implements the method described with reference to FIG. 3.

[0084] A person skilled in the art should be aware that in the above-mentioned one or more examples, functions described in the present specification can be implemented by hardware, software, firmware, or any combination thereof. When the present specification is implemented by software, the functions can be stored in a computer readable medium or transmitted as one or more instructions or code in the computer readable medium.

[0085] The objectives, technical solutions, and beneficial effects of the present specification are further described in detail in the above-mentioned specific implementations. It should be understood that the above-mentioned descriptions are merely specific implementations of the present specification, but are not intended to limit the protection scope of the present specification. Any modification, equivalent replacement, or improvement made based on the technical solutions of the present specification shall fall within the protection scope of the present specification.

CLAIMS

What is claimed is:

1. A marketing method based on a deep reinforcement learning system, wherein the deep reinforcement learning system comprises an agent and an execution environment, the agent is configured to determine a marketing behavior based on a marketing strategy and status information of the execution environment, and the method comprises:

obtaining, from the execution environment, multiple execution results generated when a user responds to the marketing behavior, wherein the multiple execution results respectively correspond to multiple target effects on a marketing effect chain;

determining a reward score of reinforcement learning based on the multiple execution results by:

determining at least one independent function f_i and least one combination function F_i, wherein the independent function f comprises at least one execution result (X_i) selected from the multiple execution results, and the combination function F comprises at least two execution results (X_i, X_{i+1}) selected from the multiple execution results;

determining a target effect Y_i for each execution result X_i, wherein the execution result X_i is designed to be closer to the target effect Y_i, and wherein the independent function f_i and the combination function F_i are designed to output a higher score when the execution result X_i is closer to the target effect Y_i;

determining the reward score based on a score evaluation function:

$$Rt=W_1\times f_1(X_1)+W_2F_2\times F_2(X_2, X_3)\ldots$$

wherein Rt is the reward score and W_i is a weight predetermined corresponding to each execution result X_i; and

returning the reward score to the agent for the agent to update its marketing strategy.

2. The method according to claim 1, wherein the obtaining, from the execution environment, multiple execution results generated when a user responds to the marketing behavior comprises:

obtaining an interaction behavior of the user from a server, and obtaining the multiple execution results by using the interaction behavior; or

capturing an operation behavior of the user by using a client device user interface

tracing point, and obtaining the multiple execution results by using the operation behavior of the user.

3. The method according to claim 1, wherein the multiple target effects on the marketing effect chain comprise multiple targets in the following: reach, a first click, stay, a further click, and conversion of marketing information.

4. The method according to claim 1, wherein the determining a reward score of reinforcement learning comprises determining the reward score by using the score evaluation function, wherein the score evaluation function uses the multiple execution results as variables and is designed to be positively correlated with a match rate between the multiple execution results and multiple corresponding target effects.

5. The method according to claim 1, wherein the determining a reward score of reinforcement learning comprises:

obtaining independent functions constructed separately for at least one of the multiple execution results, and determining the reward score based on at least a weighted sum of the independent functions.

6. The method according to claim 5, wherein at least one of the independent functions is constructed to be positively correlated with a corresponding match rate between the execution result and the target effect.

7. The method according to claim 5, wherein the multiple execution results comprise a first execution result, and the first execution result has a discrete result value; and the independent functions comprise a first function, and the first function outputs a discrete function value based on the discrete result value of the first execution result.

8. The method according to claim 7, wherein the first function outputs a first discrete value when the first execution result has a first result value and outputs a second discrete value when the first execution result has a second result value, wherein the first result value has a higher match rate with a first target effect corresponding to the first execution result than the second result value, and the first discrete value is greater than the second discrete value.

9. The method according to claim 1 or any one of claims 5 to 8, wherein the determining a reward score of reinforcement learning comprises:

determining at least one result combination based on the multiple execution results;

obtaining the at least one combination function constructed for the at least one result combination; and

determining the reward score based on at least a weighted sum of the at least one

combination function.

10. The method according to claim 1, wherein the determining a reward score of reinforcement learning comprises:

determining the target effect or a target effect combination that matches the multiple execution results;

obtaining a pre-established mapping relationship between the target effect or a target effect combination and the reward score; and

determining, based on the mapping relationship, a reward score corresponding to the matched target effect or target effect combination.

11. A marketing apparatus based on a deep reinforcement learning system, wherein the deep reinforcement learning system comprises an agent and an execution environment, the agent is configured to determine a marketing behavior based on a marketing strategy and status information of the execution environment, and the apparatus comprises:

an acquisition unit, configured to obtain, from the execution environment, multiple execution results generated when a user responds to the marketing behavior, wherein the multiple execution results respectively correspond to multiple target effects on a marketing effect chain;

a determining unit, configured to determine a reward score of reinforcement learning based on the multiple execution results by:

determining at least one independent function f_i and least one combination function F_i, wherein the independent function f comprises at least one execution result (X_i) selected from the multiple execution results, and the combination function F comprises at least two execution results (X_i, X_{i+1}) selected from the multiple execution results;

determining a target effect Y_i for each execution result X_i, wherein the execution result X_i is designed to be closer to the target effect Y_i, and wherein the independent function f_i and the combination function F_i are designed to output a higher score when the execution result X_i is closer to the target effect Y_i;

determining the reward score based on a score evaluation function:

$$Rt=W_1\times f_1(X_1)+W_2F_2\times F_2(X_2, X_3)\ldots$$

wherein Rt is the reward score and W_i is a weight predetermined corresponding to each execution result X_i; and

a returning unit, configured to return the reward score to the agent for the agent to

update its marketing strategy.

12. The apparatus according to claim 11, wherein the acquisition unit is configured to:

obtain an interaction behavior of the user from a server, and obtain the multiple execution results by using the interaction behavior; or

capture an operation behavior of the user by using a client device user interface tracing point, and obtain the multiple execution results by using the operation behavior of the user.

13. The apparatus according to claim 11, wherein the multiple target effects on the marketing effect chain comprise multiple targets in the following: reach, a first click, stay, a further click, and conversion of marketing information.

14. The apparatus according to claim 11, wherein the determining unit is configured to determine the reward score by using the score evaluation function, wherein the score evaluation function uses the multiple execution results as variables and is designed to be positively correlated with a match rate between the multiple execution results and multiple corresponding target effects.

15. The apparatus according to claim 11, wherein the determining unit is configured to:

obtain independent functions constructed separately for at least one of the multiple execution results, and determine the reward score based on at least a weighted sum of the independent functions.

16. The apparatus according to claim 15, wherein at least one of the independent functions is constructed to be positively correlated with a corresponding match rate between the execution result and the target effect.

17. The apparatus according to claim 15, wherein the multiple execution results comprise a first execution result, and the first execution result has a discrete result value; and the independent functions comprise a first function, and the first function outputs a discrete function value based on the discrete result value of the first execution result.

18. The apparatus according to claim 17, wherein the first function outputs a first discrete value when the first execution result has a first result value and outputs a second discrete value when the first execution result has a second result value, wherein the first result value has a higher match rate with a first target effect corresponding to the first execution result than the second result value, and the first discrete value is greater than the second discrete value.

19. The apparatus according to claim 11 or any one of claims 15 to 18, wherein the determining unit is configured to:

determine at least one result combination based on the multiple execution results;

obtain the at least one combination function constructed for the at least one result combination; and

determine the reward score based on at least a weighted sum of the at least one combination function.

20. The apparatus according to claim 11, wherein the determining unit is configured to:

determine the target effect or a target effect combination that matches the multiple execution results;

obtain a pre-established mapping relationship between the target effect or a target effect combination and the reward score; and

determine, based on the mapping relationship, a reward score corresponding to the matched target effect or target effect combination.

21. A computing device, comprising a memory and a processor, wherein the memory stores executable code, and when executing the executable code, the processor implements the method according to any one of claims 1 to 10.