

METHOD AND APPARATUS FOR DISPLAYING PAGE TO USER THROUGH REINFORCEMENT LEARNING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to Chinese Patent Application No. 201910127332.9, filed on February 20, 2019, the entire content of which is incorporated herein by reference.

TECHNICAL FIELD

[0002] One or more implementations of this specification relate to the field of machine learning, and in particular, to a method and an apparatus for displaying a page to a user through reinforcement learning.

BACKGROUND

[0003] In the information age, various information is displayed to users through pages. For example, various information can be displayed to users through web pages, and various APPs are also displayed to users through application pages, for example, a merchandise recommendation page of mobile phone client software of TAOBAO. How to display pages to best satisfy users' demands is an important problem.

[0004] Usually, a dedicated product manager or designer designs a page, and determines a method of displaying the page. However, this solution relies heavily on the experience of the product manager or designer, has high labor costs, and may deviate from user's demands in product design. In addition, the A/B test is sometimes used to determine a display solution. However, as such, a page display solution favored by most people can be selected, which cannot show a diversity of display methods, and cannot satisfy demands of different users to a maximum extent.

[0005] Therefore, an improved solution is desired for determining the most suitable page display methods for different users to satisfy users' demands and improve user experience.

SUMMARY

[0006] One or more implementations of this specification describe a method and an apparatus for displaying a page to a user. According to the method in the implementations, a corresponding page layout is determined based on user features of a user by using a display policy through reinforcement learning. As such, a suitable page display method is provided for the user.

[0007] According to a first aspect, a method for displaying a page to a user through reinforcement learning is provided, including the following: obtaining user features of a first user in response to a request for accessing a first page by the first user, where the user features include one or more user profile features and one or more user context features, mapping, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page, displaying the first page to the first user with the first page layout, obtaining operation information of the first user on the first page with the first page layout, and determining a reward score for current page display based on the operation information, and updating the display policy for displaying pages by using a reinforcement learning algorithm based at least on the reward score.

[0008] In an implementation, the user profile features include at least one of the following: age, gender, registration duration, education level, occupation, and user tag, and the user context features include at least one of the following: time when the request is sent, an interface that the request is sent through, or operation track information of the user.

[0009] According to an implementation, the previous first plurality of candidate page layouts are generated in advance in the following way: determining candidate page display methods, determining a candidate method of arranging page elements on the first page, determining candidate attributes of the page elements, and combining the page display methods, the candidate arrangement method, and the candidate attributes to generate the first plurality of candidate page layouts.

[0010] More specifically, in an implementation, the candidate page display methods include waterfall display and page turning display, and the candidate attributes include display sizes, display shapes, and resolution of the page elements.

[0011] In an implementation, the display policy includes an initial policy, and correspondingly, the mapping the user features to a first page layout in a plurality of page layouts that is generated in advance for the first page includes the following: randomly

assigning a first traffic bucket in a first plurality of traffic buckets to the first user, where the first plurality of traffic buckets correspond to the first plurality of candidate page layouts, and using a candidate page layout corresponding to the first traffic bucket as the first page layout.

[0012] According to different implementations, the display policy can be implemented by one of a logistic regression model, a gradient boosting decision tree GBDT model, and a deep neural network.

[0013] In an implementation, the previous operation information includes at least one of the following: hit count on the page, page view, view depth to the page, and time on the page, and correspondingly, the reward score can be determined as a weighted sum of at least one of the previous operation information.

[0014] In an implementation, the operation information includes a transaction amount achieved based on the first page, and correspondingly, the reward score can be determined to be directly proportional to the transaction amount.

[0015] According to an implementation, the updating the display policy for displaying pages by using a reinforcement learning algorithm based at least on the reward score includes the following: adding the user features, the first page layout, and the reward score to a training sample set as one training sample, and updating the display policy for displaying pages by using the training sample set through the reinforcement learning algorithm.

[0016] In different implementations, the reinforcement learning algorithm can be one of the following: an evolution strategy algorithm, a Q-learning algorithm, a SARSA algorithm, a policy gradient algorithm, and a cross information entropy method.

[0017] According to a second aspect, an apparatus for displaying a page to a user through reinforcement learning is provided, including the following: a feature acquisition unit, configured to obtain user features of a first user in response to a request for accessing a first page by the first user, where the user features include one or more user profile features and one or more user context features, a layout determining unit, configured to map, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page, a display unit, configured to display the first page to the first user with the first page layout, a feedback acquisition unit, configured to obtain operation information of the first user on the first page with the first page layout, and determine a reward score for current page display based on the operation information, and an update unit, configured to update the display policy by using a reinforcement learning algorithm based at least on the reward score to perform page

display.

[0018] According to a third aspect, a computer-readable storage medium is provided, a computer program is stored on the computer-readable storage medium, and when the computer program is executed in a computer, the computer performs the method in the first aspect.

[0019] According to a fourth aspect, a computing device is provided, including a memory and a processor, where the memory stores executable code, and when executing the executable code, the processor performs the method in the first aspect.

[0020] According to the method and the apparatus provided in the implementations of this specification, a page display policy can be learned through policy exploration and policy update in reinforcement learning without labeling. In the page display policy, the most suitable page display methods can be determined for different users. As such, a diversity of page display methods are shown, the personalized demands of users are satisfied, and the use experience of users is improved.

BRIEF DESCRIPTION OF DRAWINGS

[0021] To describe the technical solutions in the implementations of the present disclosure more clearly, the following briefly describes the accompanying drawings for describing the implementations. Apparently, the accompanying drawings in the following description show merely some implementations of the present disclosure, and a person of ordinary skill in the art can derive other drawings from these accompanying drawings without creative efforts.

[0022] FIG. 1 is a schematic diagram illustrating a typical reinforcement learning system;

[0023] FIG. 2 is a schematic diagram illustrating an implementation scenario of an implementation disclosed in this specification;

[0024] FIG. 3 is a flowchart illustrating a method for displaying a page to a user, according to an implementation; and

[0025] FIG. 4 is a schematic diagram illustrating an apparatus, according to an implementation.

DESCRIPTION OF IMPLEMENTATIONS

[0026] The following describes the solution provided in this specification with

reference to the accompanying drawings.

[0027] To perform personalized and targeted page display for different users, according to the implementations of this specification, the concept and the method of reinforcement learning are used to determine suitable page display solutions for the users. As known by a person skilled in the art, reinforcement learning is an unlabeled learning method that is based on sequence behavior feedbacks and is policy learning performed through continuous "trials and errors".

[0028] FIG. 1 is a schematic diagram illustrating a typical reinforcement learning system. As shown in FIG. 1, usually, the reinforcement learning system includes an agent and an execution environment. The agent continuously learns and improves a policy of the agent by interacting with and feeding back to the execution environment. Specifically, the agent observes and obtains a state (state) of the execution environment, and determines, based on a specific policy, a behavior or an action (action) to be taken for the state of the current execution environment. When such a behavior acts on the execution environment, the state of the execution environment can be changed, and a feedback is generated to the agent at the same time. The feedback is also referred to as reward score (reward). The agent determines, based on the obtained reward score, whether the previous behavior is correct and whether the policy needs to be adjusted to further update the policy of the agent. By repeatedly observing states, determining actions, and receiving feedbacks, the agent can repeatedly update the policy, and the ultimate objective of the agent is to learn a policy so that the obtained cumulative reward score is maximized.

[0029] FIG. 2 is a schematic diagram illustrating an implementation scenario of an implementation disclosed in this specification. According to one or more implementations of this specification, in the concept of reinforcement learning, when a user requests to access a page, a related feature of the user is extracted (which is equivalent to obtaining state s of an execution environment), and a specific display policy is used to determine a page layout suitable for the user for display (which is equivalent to determining action a to be taken). Then, the interaction and operation information of the user on the page is used as reward score r for feedback, and is used as a learning sample to update a display policy. In such repeated learning, the display policy is improved by repeated explorations and "trials and errors". Therefore, the improved display policy can be used to determine the most suitable page display methods for different users.

[0030] The following describes a specific implementation method of the previous concept.

[0031] FIG. 3 is a flowchart illustrating a method for displaying a page to a user,

according to an implementation. It can be understood that the method can be performed by any apparatus, device, platform, or a device cluster that has computing and processing capabilities. As shown in FIG. 3, the method includes the following: Step 31: Obtain user features of a first user in response to a request for accessing a first page by the first user, where the user features include one or more user profile features and one or more user context features. Step 32: Map, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page. Step 33: Display the first page to the first user with the first page layout. Step 34: Obtain operation information of the first user on the first page with the first page layout, and determine a reward score for current page display based on the operation information. Step 35: Update the display policy by using a reinforcement learning algorithm based at least on the reward score to perform page display. The following describes specific execution of the previous steps.

[0032] First, in step 31, the user features of the first user are obtained in response to the request for accessing the first page by the first user. It should be understood that the first user here means any specific user. The first page can be any page. For example, the first page can be a web page, or can be an APP page. More specifically, in an example, the first page is the homepage or the merchandise recommendation page of the TMALL APP.

[0033] As such, when the first user requests to access the first page, the user features of the first user are obtained. The user features can include user profile features. In an implementation, the user profile features can include basic attribute features of the user, such as age, gender, registration duration, education level, and occupation.

[0034] In an implementation, the user profile features can further include a user tag that can directly reflect a feature of the user. Usually, the user tag can be determined based on user attribute features and historical big data of users. There are a plurality of solutions of classifying users as crowds. In these solutions, user tags are added to users in various ways such as clustering and grouping, for example, "workplace newbie" and "full-time housewife". In step 31, various known crowd classification solutions can be used to obtain the previous user tag. Alternatively, tags can be added to users by using various crowd classification solutions in advance. In step 31, a user tag of the user is read.

[0035] The user features can further include user context features. In an implementation, the user context features can include time when the page request is sent, such as the morning, the working time, or the evening. The user context features can further include an interface that the page request is sent through, for example, a page or a

link used by the user to send the current page request. The user context features can further include operation track information of the user, that is, operations that are performed by the user before the page request is sent, for example, historical pages that are accessed, or operations (click, zoom-in, zoom-out, and input) that are performed on the historical pages.

[0036] The previously obtained user features are equivalent to current state s of an execution environment in the reinforcement learning system. Then, action a to be taken in current state s needs to be determined based on a policy. Therefore, in step 32, the previous user features are mapped based on the display policy to the first page layout in N candidate page layouts generated for the first page.

[0037] The previous N candidate page layouts for the first page can be generated in advance in a plurality of ways.

[0038] In an implementation, candidate page display methods are determined for the first page in advance. Specifically, for the APP page, the candidate page display methods can include waterfall display and page turning display. For the web page, the candidate page display methods can include single-page display, drop-down display, etc.

[0039] In addition, a candidate method of arranging page elements on the first page is further determined. For example, the page elements are displayed in several rows, and several elements are displayed in each row. Candidate attributes of the page elements can be further determined. These candidate attributes include display sizes, display shapes, resolution, etc. of the page elements.

[0040] A plurality of candidate page layouts can be generated by combining the previous page display methods, the candidate arrangement method of the page elements, and the candidate attributes.

[0041] In addition, in some implementations, more factors related to page display can be further selected, such as a background color of the page, a font style of the page, and screen brightness of the page. By further selecting the previous factors, further modification and adjustment are performed on the page layout to obtain more candidate page layouts.

[0042] More specifically, assume that the first page is a merchandise recommendation page of an e-commerce APP (for example, TMALL in a mobile phone), and page elements on the first page are merchandises to be displayed. Then, two different page display methods (waterfall and page turning display), m different merchandise arrangement ways (for example, four merchandises in each row or three merchandises in each row), and k merchandise attribute selections (for example, sizes, shapes, and resolution of a

merchandise picture) can be combined to obtain $N=2*m*k$ candidate page layouts.

[0043] In another implementation, a designer can design a plurality of candidate page layouts in advance based on different crowd and different styles, as the previous N candidate page layouts.

[0044] The previous N candidate page layouts generated in advance are equivalent to N candidate actions, that is, N-dimensional action space is generated. Correspondingly, the display policy in step 32 is used to select an appropriate action, that is, an appropriate page layout from the N-dimensional action space based on the user features. In other words, the display policy is equivalent to a mapping function, and the mapping function is used to map the user features to one of the N candidate page layouts, which is referred to as a first page layout below.

[0045] In an implementation, the previous mapping function can be implemented by using an N-classification algorithm, and each classification corresponds to one candidate page layout. As such, in the display policy, the user is classified into one of the N classifications based on the previous user features, and therefore a corresponding candidate page layout is determined.

[0046] According to different implementations, the previous display policy can be implemented by using various machine learning models, for example, a logistic regression model, a gradient boosting decision tree GBDT model, and a deep neural network.

[0047] Initially, the display policy can be a random policy, and a current candidate page layout suitable for the first user is determined randomly. Specifically, in an implementation, online traffic is divided into N traffic buckets, and the N traffic buckets correspond to N candidate page layouts. Therefore, a certain traffic bucket in the N traffic buckets can be randomly assigned to the current first user and is denoted as a first traffic bucket. Correspondingly, a candidate page layout corresponding to the first traffic bucket can be used as the first page layout selected for the first user.

[0048] Certainly, the previous random policy is a temporary policy in initial "trial and error" and policy exploration processes. The display policy is continuously improved, updated, and adjusted through user feedbacks, which is further described later.

[0049] After the corresponding first page layout is determined for the first user in step 32, the first page is displayed to the first user with the first page layout in step 33. This is equivalent to sending execution action a to the execution environment.

[0050] Then, in step 34, the operation information of the first user on the first page with the first page layout is obtained, and a reward score for current page display is

determined based on the operation information.

[0051] It can be understood that the reward score in reinforcement learning is a measurement of whether the previously performed action a is correct, and is also a main basis for subsequent policy adjustment. The objective of the policy adjustment is to maximize the cumulative reward scores obtained by performing action a . In a page display scenario, the objective of the policy training is to enable the page layout determined by the display policy to achieve a maximum user satisfaction degree, and/or enable the page platform to obtain maximum benefits by using the page. Therefore, the reward score can be set to represent, by using an operation behavior of the user on the page, the satisfaction degree of the user on the page and/or the benefits of the page platform.

[0052] Specifically, after the first page of the first page layout is returned to the first user, in an implementation, one or more of the following operation information of the first user on the first page can be obtained: hit count a_1 on the page, view depth a_2 to the page, page view a_3 , time a_4 on the page, etc. It can be assumed that the previous operation information has a positive correlation with the satisfaction degree of the user, because when the user is not satisfied with the page layout of the first page, the user tends to leave the page or abandon the operation.

[0053] Therefore, in an implementation, the reward score for current page display is determined as a weighted sum of at least one of the previous operation information. For example, the reward score is determined as:

$$R = w_1 * a_1 + w_4 * a_4 \quad (1)$$

w_1 and w_4 are weight factors.

[0054] In an implementation, the previous first page is a page related to a transaction, such as an e-commerce APP page. In this case, the following operation information of the user can be obtained: transaction amount TX achieved based on the first page, for example, a purchase amount. On this basis, reward score R for current page display can be determined to be directly proportional to transaction amount TX , that is,

$$R = W * TX \quad (2)$$

[0055] In an implementation, ways of determining the previous reward score can be further combined to obtain more specific representation ways of the reward score.

[0056] For example, the reward score can be determined as:

$$R = w_1 * a_1 + W * TX \quad (3)$$

[0057] That is, the reward score is a weighted sum of hit count a_1 of the user on the first page and the transaction amount.

[0058] As such, operation information of the user on the first page is obtained, and a reward score is determined based on the operation information of the user as feedback information for current page display.

[0059] Then, in step 35, the display policy is updated by using a reinforcement learning algorithm based at least on the reward score to perform page display. According to the concept of policy update in reinforcement learning, in this implementation of this specification, the display policy is repeatedly updated based on feedback information (that is, a reward score), and the objective of the update is to enable the page layout (action a) determined by the display policy for the same user feature (environment state s) to obtain more expected cumulative reward scores.

[0060] To achieve such an objective, in an implementation, a policy update is performed every time a reward score of a user is obtained. For example, a SARSA algorithm can be used to directly perform an action selection based on the current display policy, and then the current action selection is used as a sample to update the current policy.

[0061] In a more common case, feedback information of a plurality of users for a plurality of page layouts is collected as a training sample set, and a display policy is trained by using the training sample set to update the display policy.

[0062] Specifically, for the previous first user, after the first user performs an operation on the first page of the first page layout to obtain a reward score for current page display, the user features (equivalent to environment state s) of the first user, the first page layout (equivalent to action a), and the reward score (r) are added to the training sample set as one training sample. As such, the training sample set includes a plurality of training samples in a form of $\langle s, a, r \rangle$. In an implementation, in a continuous-interaction scenario, a new user feature (where the user profile feature does not change, but the context feature can change) is further obtained after the first user performs an operation on the first page layout. This is equivalent to obtaining new state s' that an execution environment is transferred into after action a is performed. As such, a training sample in the form of $\langle s, a, r, s' \rangle$ is generated.

[0063] Then, the display policy is updated by using the training sample set through the reinforcement learning algorithm. According to different implementations, a plurality of reinforcement learning algorithms can be used for training. These reinforcement learning algorithms include an evolution strategy algorithm, a Q-learning algorithm, a policy gradient algorithm, a cross information entropy method, etc.

[0064] The following describes an update process of a display policy with reference

to a typical Q-learning algorithm.

[0065] The update of the display policy can include evaluation of the policy and improvement of the policy. In the Q-learning algorithm, the state-action value function (state-action value function) Q is used to represent an expected cumulative reward. The state-action value function is also referred to as a Q function, where Q(s, a) represents the cumulative reward score brought by using policy π after action a is performed from state s. The value of the Q function is referred to as a Q value. The cumulative reward score can be calculated in the form of T-step cumulative reward or γ -discount cumulative reward. The expressions of the Q function in these two forms are as follows:

$$\begin{cases} Q_T^\pi(s, a) = E_\pi \left[\frac{1}{T} \sum_{t=1}^T r_t \right] \\ Q_\gamma^\pi(s, a) = E_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_{t+1} \right] \end{cases}$$

[0066] The subscript t represents a quantity of subsequent steps to be performed, and r_t is a reward score for t steps.

[0067] As described above, a plurality of training samples in the form of <s, a, r, s'> are collected. By using these training samples, reward scores r obtained after different actions a are applied in different states s can be determined. Further, it is learned that Q values are correspondingly obtained by applying different actions a in different states s.

[0068] During a policy update, if a greedy algorithm is used, the following display policy can be obtained: For state s formed by current user features, respective Q values for displaying N candidate page layouts (that is, N different actions are applied) in state s are determined, and a page layout corresponding to a maximum Q value in the N Q values is selected as a display layout.

[0069] A policy gradient algorithm uses different policy evaluation and update ways. In the algorithm, learning and evaluation of Q functions are not involved. Instead, a gradient of policy parameter adjustment is directly determined for a display policy with smaller loss, and the policy is adjusted based on the determined gradient.

[0070] Other reinforcement learning algorithms are not described one by one here. By using these algorithms, a display policy for a page can be updated. A direction of the update is to enable the displayed page layout to obtain more cumulative reward scores, that is, obtain a greater user satisfaction degree. The display policy obtained through this update can be used to determine the most suitable page layout for different users.

[0071] In conclusion, according to the solutions of the previous implementations, a

page display policy can be learned through policy exploration and policy update in reinforcement learning without labeling. In the page display policy, the most suitable page display methods can be determined for different users. As such, a diversity of page display methods are shown, the personalized demands of users are satisfied, and the use experience of users is improved.

[0072] According to an implementation of another aspect, an apparatus for displaying a page to a user through reinforcement learning is further provided. The apparatus can be deployed at any device or platform that has computing and processing capabilities. FIG. 4 is a schematic diagram illustrating an apparatus, according to an implementation. As shown in FIG. 4, the apparatus 400 includes the following: a feature acquisition unit 41, configured to obtain user features of a first user in response to a request for accessing a first page by the first user, where the user features include one or more user profile features and one or more user context features, a layout determining unit 42, configured to map, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page, a display unit 43, configured to display the first page to the first user with the first page layout, a feedback acquisition unit 44, configured to obtain operation information of the first user on the first page with the first page layout, and determine a reward score for current page display based on the operation information, and an update unit 45, configured to update the display policy by using a reinforcement learning algorithm based at least on the reward score to perform page display.

[0073] In an implementation, the previous user profile features include at least one of the following: age, gender, registration duration, education level, occupation, and user tag, and the user context features include at least one of the following: time when the request is sent, an interface that the request is sent through, or operation track information of the user.

[0074] According to an implementation, the previous first plurality of candidate page layouts are generated in advance in the following way: determining candidate page display methods, determining a candidate method of arranging page elements on the first page, determining candidate attributes of the page elements, and combining the page display methods, the candidate arrangement method, and the candidate attributes to generate the first plurality of candidate page layouts.

[0075] More specifically, in an implementation, the previous candidate page display methods include waterfall display and page turning display, and the previous candidate attributes include display sizes, display shapes, and resolution of the page elements.

[0076] In an implementation, the previous display policy can be an initial policy, and correspondingly, the layout determining unit 42 is configured to: randomly assign a first traffic bucket in a first plurality of traffic buckets to the first user, where the first plurality of traffic buckets correspond to the first plurality of candidate page layouts, and use a candidate page layout corresponding to the first traffic bucket as the first page layout.

[0077] In different implementations, the display policy is implemented by one of a logistic regression model, a gradient boosting decision tree GBDT model, and a deep neural network.

[0078] According to an implementation, the operation information can include at least one of the following: hit count on the page, page view, view depth to the page, and time on the page.

[0079] Correspondingly, the feedback acquisition unit 44 is configured to determine the reward score as a weighted sum of at least one of the operation information.

[0080] According to an implementation, the operation information can include a transaction amount achieved based on the first page.

[0081] Correspondingly, the feedback acquisition unit 44 can be configured to determine the reward score to be directly proportional to the transaction amount.

[0082] In an implementation, the update unit 45 is configured to: add the user features, the first page layout, and the reward score to a training sample set as one training sample and update the display policy by using the training sample set through the reinforcement learning algorithm.

[0083] According to different implementations, the reinforcement learning algorithm can include an evolution strategy algorithm, a Q-learning algorithm, a SARSA algorithm, a policy gradient algorithm, and a cross information entropy method.

[0084] The previous apparatus can provide respective suitable page layouts for different users, thereby improving user experience.

[0085] According to an implementation of another aspect, a computer-readable storage medium is further provided, a computer program is stored on the computer-readable storage medium, and when the computer program is executed in a computer, the computer performs the method described with reference to FIG. 3.

[0086] According to an implementation of the 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.

[0087] A person skilled in the art should be aware that in the previous one or more

examples, functions described in the present disclosure can be implemented by hardware, software, firmware, or any combination thereof. When implemented by software, the previous functions can be stored in a computer-readable medium or transmitted as one or more instructions or code in the computer-readable medium.

[0088] The objectives, technical solutions, and benefits of the present disclosure are further described in detail in the previous specific implementations. It should be understood that the previous descriptions are merely specific implementations of the present disclosure, but are not intended to limit the protection scope of the present disclosure. Any modification, equivalent replacement, or improvement made based on the technical solutions of the present disclosure shall fall within the protection scope of the present disclosure.

CLAIMS

What is claimed is:

1. A method for displaying a page to a user through reinforcement learning, comprising:

obtaining user features of a first user in response to a request for accessing a first page by the first user, wherein the user features comprise one or more user profile features and one or more user context features;

mapping, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page;

displaying the first page to the first user with the first page layout;

obtaining operation information of the first user on the first page with the first page layout, and determining a reward score for current page display based on the operation information; and

updating the display policy for displaying pages by using a reinforcement learning algorithm based at least on the reward score to perform page display.

2. The method according to claim 1, wherein

the user profile features comprise at least one of the following: an age, a gender, registration duration, an education level, an occupation, and a user tag; and

the user context features comprise at least one of the following: time when the request is sent, an interface that the request is sent through, or operation track information of the user.

3. The method according to claim 1, wherein the first plurality of candidate page layouts are generated in advance in the following way:

determining candidate page display methods;

determining a candidate method of arranging page elements on the first page;

determining candidate attributes of the page elements; and

combining the page display methods, the candidate arrangement method, and the candidate attributes to generate the first plurality of candidate page layouts.

4. The method according to claim 3, wherein

the candidate page display methods comprise waterfall display and page turning display; and

the candidate attributes comprise display sizes, display shapes, and resolution of the page elements.

5. The method according to claim 1, wherein the display policy is an initial policy,

and the mapping the user features to a first page layout in a plurality of page layouts that is generated in advance for the first page comprises:

randomly assigning a first traffic bucket in a first plurality of traffic buckets to the first user, wherein the first plurality of traffic buckets correspond to the first plurality of candidate page layouts; and

using a candidate page layout corresponding to the first traffic bucket as the first page layout.

6. The method according to claim 1, wherein the display policy is implemented by one of a logistic regression model, a gradient boosting decision tree GBDT model, and a deep neural network.

7. The method according to claim 1, wherein the operation information comprises at least one of the following: hit count on the page, page view, view depth to the page, and time on the page; and

the determining a reward score for current page display based on the operation information comprises: determining the reward score as a weighted sum of at least one of the operation information.

8. The method according to claim 1 or 7, wherein the operation information comprises a transaction amount achieved based on the first page; and

the determining a reward score for current page display based on the operation information comprises: determining the reward score to be directly proportional to the transaction amount.

9. The method according to claim 1, wherein the updating the display policy for displaying pages by using a reinforcement learning algorithm based at least on the reward score to perform page display comprises:

adding the user features, the first page layout, and the reward score to a training sample set as one training sample; and

updating the display policy for displaying pages by using the training sample set through the reinforcement learning algorithm.

10. The method according to claim 1 or 9, wherein the reinforcement learning algorithm comprises a SARSA algorithm, an evolution strategy algorithm, a Q-learning algorithm, a policy gradient algorithm, and a cross information entropy method.

11. An apparatus for displaying a page to a user through reinforcement learning, comprising:

a feature acquisition unit, configured to obtain user features of a first user in response to a request for accessing a first page by the first user, wherein the user features comprise

one or more user profile features and one or more user context features;

a layout determining unit, configured to map, based on a display policy, the user features to a first page layout in a first plurality of candidate page layouts generated for the first page;

a display unit, configured to display the first page to the first user with the first page layout;

a feedback acquisition unit, configured to obtain operation information of the first user on the first page with the first page layout, and determine a reward score for current page display based on the operation information; and

an update unit, configured to update the display policy by using a reinforcement learning algorithm based at least on the reward score to perform page display.

12. The apparatus according to claim 11, wherein

the user profile features comprise at least one of the following: an age, a gender, registration duration, an education level, an occupation, and a user tag; and

the user context features comprise at least one of the following: time when the request is sent, an interface that the request is sent through, or operation track information of the user.

13. The apparatus according to claim 11, wherein the first plurality of candidate page layouts are generated in advance in the following way:

determining candidate page display methods;

determining a candidate method of arranging page elements on the first page;

determining candidate attributes of the page elements; and

combining the page display methods, the candidate arrangement method, and the candidate attributes to generate the first plurality of candidate page layouts.

14. The apparatus according to claim 13, wherein

the candidate page display methods comprise waterfall display and page turning display; and

the candidate attributes comprise display sizes, display shapes, and resolution of the page elements.

15. The apparatus according to claim 11, wherein the display policy is an initial policy, and the layout determining unit is configured to:

randomly assign a first traffic bucket in a first plurality of traffic buckets to the first user, wherein the first plurality of traffic buckets correspond to the first plurality of candidate page layouts; and

use a candidate page layout corresponding to the first traffic bucket as the first page

lay out.

16. The apparatus according to claim 11, wherein the display policy is implemented by one of a logistic regression model, a gradient boosting decision tree GBDT model, and a deep neural network.

17. The apparatus according to claim 11, wherein the operation information comprises at least one of the following: hit count on the page, page view, view depth to the page, and time on the page; and

the feedback acquisition unit is configured to determine the reward score as a weighted sum of at least one of the operation information.

18. The apparatus according to claim 11 or 17, wherein the operation information comprises a transaction amount achieved based on the first page; and

the feedback acquisition unit is configured to determine the reward score to be directly proportional to the transaction amount.

19. The apparatus according to claim 11, wherein the update unit is configured to:
add the user features, the first page layout, and the reward score to a training sample set as one training sample; and

update the display policy by using the training sample set through the reinforcement learning algorithm.

20. The apparatus according to claim 11 or 19, wherein the reinforcement learning algorithm comprises an evolution strategy algorithm, a Q-learning algorithm, a SARSA algorithm, a policy gradient algorithm, and a cross information entropy method.

21. A computer-readable storage medium, wherein a computer program is stored on the computer-readable storage medium, and when the computer program is executed in a computer, the computer performs the method according to any one of claims of 1 to 10.

22. 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.