

DIGITAL AVATAR RECOMMENDATION METHOD AND RECOMMENDATION SYSTEM

TECHNICAL FIELD

[0001] One or more implementations of the present specification relate to the field of machine
5 learning, and in particular, to a method and system for performing recommendation by using a
digital avatar.

BACKGROUND

[0002] On the Internet environment, there are various recommendation scenarios, such as
shopping recommendation, video recommendation generated during short video watching, and
10 friend recommendation in social networks. Therefore, a recommendation system needs to be used
to provide individualized recommendation for a user, to display content that is more likely to be
of interest to the user, including commodities, articles, and other objects, and better provide
targeted services for the user.

[0003] Generally, the recommendation system typically uses the following working mode:
15 First, a batch of candidate content items that are possibly of interest to the user are prepared, and
then the candidate content items are ranked based on a feature of the user, and several content
items that rank in the front are recalled. Finally, the several recalled content items are displayed
on a page. In the above mode, due to a limitation of a display location on the display page, only a
small quantity of content items can be displayed, and a large quantity of content items have no
20 chance to be displayed. In another aspect, recommendation can only be passively displayed and
pushed to the user, and generally, the user cannot perform continuous interaction, which is not
conducive to understand a preference and a requirement of the user.

[0004] Therefore, the recommendation effect of the conventional recommendation system
needs to be further improved.

SUMMARY

[0005] One or more implementations of the present specification describe a digital avatar recommendation method and system, and a recommendation decision can be made by using a digital avatar through a reinforcement learning policy, so as to implement a better recommendation effect and improve user experience.

[0006] According to a first aspect, a recommendation method is provided and is performed by a digital avatar recommendation system, the digital avatar recommendation system includes a computer-simulated digital avatar, and the method includes: obtaining current state data, where the state data includes user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar; mapping, by an agent in the digital avatar, the state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, where a candidate action in the candidate action set corresponds to a to-be-recommended content category, and the target action corresponds to a target content category; and performing, by the digital avatar, target interaction with the target user, where the target interaction is used to recommend the target content category.

[0007] According to an implementation, the current scenario is a live streaming scenario, the digital avatar is a live streamer in the live streaming scenario, and the target user is a user entering the live streaming scenario.

[0008] Further, in an implementation, the scenario information includes a theme of the live streaming, category information of a store to which the live streaming belongs, and an element in the live streaming to which the target user has interacted.

[0009] According to another implementation, the scenario is a metaverse, and the digital avatar is a system role in the metaverse.

[0010] In an implementation, the method further includes: obtaining a target feedback of the target user for the target interaction; and determining, based on the target feedback, a reward score corresponding to the target action, where the reward score is used to update the current policy.

[0011] Further, the reward score corresponding to the target action can be determined in the following manner: determining that the reward score is a first value if the target feedback indicates that the target user accepts the target content category; or determining that the reward score is a second value if the target feedback indicates that the target user does not accept the target content

category, where the second value is less than the first value.

[0012] In an implementation, the mapping the state data to the target action in the candidate action set includes: determining, based on the state data, long-term cumulative reward Q values corresponding to candidate actions in the candidate action set, to obtain Q values; and determining
5 the target action based on the Q values.

[0013] According to an implementation, the state data is obtained by fusing source data of a plurality of modes, and the source data of the plurality of modes includes at least two of the following: text source data, image source data, audio source data, or video source data.

[0014] In an implementation, the user information includes relationship information obtained
10 based on a user relationship network graph, and the state data includes graph data of the relationship network graph.

[0015] In an implementation, the state data further includes graph data of a knowledge graph, and an entity in the knowledge graph corresponds to the to-be-recommended content category.

[0016] According to an implementation, the digital avatar has configuration parameters
15 defining character features of the digital avatar, and the state data further includes the configuration parameters.

[0017] According to a second aspect, a digital avatar recommendation system is provided, including a state acquisition module and a computer-simulated digital avatar, where an agent is embedded into the digital avatar.

20 [0018] The state acquisition module is configured to obtain current state data, where the state data includes user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar.

[0019] The agent is configured to map the state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, where a candidate action in the
25 candidate action set corresponds to a to-be-recommended content category, and the target action corresponds to a target content category.

[0020] The digital avatar is configured to perform target interaction with the target user, where the target interaction is used to recommend the target content category.

[0021] In an implementation, the digital avatar recommendation system further includes an
30 updating module, and the updating module is configured to: obtain a target feedback of the target user for the target interaction; determine, based on the target feedback, a reward score

corresponding to the target action; and update the current policy based on the reward score.

[0022] According to a third aspect, a computer-readable storage medium is provided. The computer-readable storage medium stores a computer program, and when the computer program is executed on a computer, the computer is enabled to perform the method according to the first aspect.

[0023] According to a fourth aspect, a computing device is provided, including a memory and a processor. The memory stores executable code, and the processor implements the method according to the first aspect when executing the executable code.

[0024] In the implementations of the present specification, a new framework in which a digital avatar performs recommendation through reinforcement learning is proposed. In the framework, the digital avatar is applied to a recommendation scenario to form a digital avatar recommendation system, so as to provide a better recommendation service by using a feature of the digital avatar. In addition, in the digital avatar recommendation system, an agent obtained through reinforcement learning is embedded or introduced into an analog program of the digital avatar to form a digital avatar recommendation engine based on reinforcement learning. Based on repeated interaction between the digital avatar and a user, the recommendation engine can better understand a requirement and a preference of the user, and more accurately perform recommendation.

BRIEF DESCRIPTION OF DRAWINGS

[0025] To describe the technical solutions in the implementations of the present specification more clearly, the following is a brief introduction of the accompanying drawings for illustrating such implementations. Clearly, the accompanying drawings described below are merely some implementations of the present specification, and a person of ordinary skill in the art can derive other drawings from such accompanying drawings without making innovative efforts.

[0026] FIG. 1 is a schematic diagram illustrating a typical reinforcement learning system;

[0027] FIG. 2 is a schematic diagram illustrating a digital avatar recommendation system according to an implementation;

[0028] FIG. 3 is an execution flowchart illustrating a recommendation method according to an implementation; and

[0029] FIG. 4 is a schematic diagram illustrating a structure of a digital avatar recommendation

system according to an implementation.

DESCRIPTION OF EMBODIMENTS

[0030] The following describes the solutions provided in the present specification with reference to the accompanying drawings.

5 **[0031]** Due to a limitation of a conventional recommendation system, in the implementations of the present specification, a digital avatar is used for recommendation to form a digital avatar recommendation system.

[0032] A digital human, also referred to as a digital avatar, is a virtual agent simulated by a computer, or is referred to as an avatar or virtual avatar. The digital avatar may have a three-
10 dimensional digital body, can communicate with a real human user by using a natural language, understand content of a statement, and infer a meaning, and can also have character features such as a facial expression, emotion, and personality feature of a real person. In a given virtual scenario, the digital avatar can further perform various body actions.

[0033] The digital avatar is formed based on a large number of machine learning models and
15 a large amount of training data, including a conversation generation model based on natural language processing NLP, a reasoning model, an action generation model, etc. In addition, a computation module, such as an animation construction module, that is configured to render and present a digital avatar form is further required. Currently, many Internet companies provide various platforms and tools, and these platforms and tools can be used to generate the digital avatar
20 or some modules of the digital avatar at low costs. Therefore, the digital avatar can be widely applied to various digital scenarios.

[0034] For example, in a rapidly growing metaverse field, the digital avatar can understand user interaction content and quickly respond to a user behavior by using an advantage of an artificial intelligence (AI) algorithm. Advantages of the digital avatar lie not only in that a
25 requirement of a user can be satisfied. In addition, because the digital avatar can have character features such as a facial expression, emotion, and personality feature that are similar to those of a real person, a response can be made in a very "humanized" friendly manner to further improve user experience.

[0035] Because of many advantages of the digital avatar, in the implementations of the present

specification, the digital avatar is applied to a recommendation system to form a digital avatar recommendation system, so as to provide a better recommendation service by using the features of the digital avatar. For example, the digital avatar can repeatedly interact with a user, to actively capture attention of the user, better understand a requirement and a preference of the user by using interaction content, and more accurately perform recommendation.

[0036] Further, to better make a recommendation decision based on an interaction between the digital avatar and the user, in the implementations of the present specification, a recommendation policy is formed through reinforcement learning.

[0037] FIG. 1 is a schematic diagram illustrating an example reinforcement learning system.

As shown in FIG. 1, generally, the reinforcement learning system includes an agent (Agent) and an execution environment. The agent constantly performs learning and optimizes a policy of the agent by interacting with and giving a feedback to the execution environment. For example, the agent observes and obtains a state (state) of the execution environment, and determines, based on a policy, a behavior or an action (action) to be taken in the current state of the execution environment. When the behavior is performed on the execution environment, the state of the execution environment is changed, and a feedback is generated to the agent. The feedback is also referred to as a reward (reward) score. The agent determines, based on the obtained reward score, whether the previous behavior is correct and whether the policy needs to be adjusted, so as to update the policy of the agent. The agent repeatedly and constantly observes the state, determines the behavior, and receives the feedback, so that the agent can constantly update the policy. An ultimate goal is to learn a policy to maximize an obtained long-term cumulative reward score.

[0038] The above process can be modeled as a Markov decision process (MDP). The MDP process can be described by using a tuple (S, A, p, r, γ) , where S represents state space that can be observed by the agent, and A represents action space that can be used by the agent, and includes all possible actions. It is assumed that in a t^{th} round of interaction, an environment state is s_t , and an action taken is a_t . In this case, the environment state is migrated to a new state s_{t+1} by using a migration probability function p , e.g., $s_{t+1} \sim p(\cdot | s_t, a_t)$. In addition, the migration generates a reward score $r_t = r(s_t, a_t)$, where γ is a discount coefficient of the reward score.

[0039] In the MDP process, the agent performs learning by repeatedly and constantly observing the state, determining the behavior, and receiving the feedback. The goal of learning is to find a policy π that can improve a given performance evaluation function $J(\pi)$. For example, the

performance evaluation function in the given policy can be defined as a discount cumulative reward of an infinite number of steps in the policy, for example:

$$J(\pi) = E_{\tau \sim \pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \right] \quad (1)$$

where $r(s_t, a_t, s_{t+1})$ indicates that in the t^{th} step, the action a_t is used in the state s_t , so that the state changes to a reward score obtained in the state s_{t+1} ; and γ is the discount coefficient.

[0040] In a value-based policy update solution, the goal of agent learning is to achieve an improved value function. The value function is a discount cumulative reward function to be achieved by executing the policy π . In an implementation, a state-action value function, e.g., a q function, can be defined to represent a long-term cumulative reward:

$$q_{\pi}(s, a) = E_{\tau \sim \pi} [r(\tau) \mid s_0 = s, a_0 = a] \quad (2)$$

[0041] The state-action value function represents a cumulative reward brought by using the policy π after an action a_0 is performed starting from a state s_0 . A cumulative reward in the t^{th} step can be obtained based on the state-action value function, which is referred to as a Q value.

[0042] The policy π adopted by the agent can be considered as a function for mapping from the state s to the action a , and a parameter in the function can be represented as θ . Therefore, the process of reinforcement learning is a process of finding out reward scores obtained by performing various actions in various states through continuous "trial and error", and continuously adjusting and updating the parameter θ based on feedbacks of the reward scores, so as to obtain a trained policy. A trained policy finally obtained can be used to perform an appropriate action a based on a current state s . This action can make the obtained long-term cumulative reward score as large as possible.

[0043] In view of the above features of reinforcement learning, the agent obtained through reinforcement learning can be embedded or introduced into an analog program of the digital avatar to form a digital avatar recommendation engine based on reinforcement learning. Based on repeated interaction between the digital avatar and a user, the recommendation engine can better understand a requirement and a preference of the user, and more accurately perform recommendation.

[0044] FIG. 2 is a schematic diagram illustrating a digital avatar recommendation system according to an implementation. As shown in FIG. 2, a digital avatar can perform a plurality of rounds of interaction with a plurality of users to form a corresponding interaction session.

According to an implementation, an agent obtained through reinforcement learning is introduced into the digital avatar. Correspondingly, different states of an execution environment can be formed based on different user information and interaction sessions, and optional action space A is formed based on a to-be-recommended content. For different users, the agent determines, based on different environment states, recommendation actions to be performed next, and the digital avatar interacts with the users based on the recommendation actions. As such, the digital avatar implements active and targeted content recommendation by interacting with the user.

[0045] The technical concept and framework of the solutions in the implementations of the present specification are described above. An example recommendation execution process of the digital avatar recommendation system is described below by combining FIG. 2 and FIG. 3.

[0046] FIG. 3 is an execution flowchart illustrating a recommendation method according to an implementation. It should be noted that the method is performed by the digital avatar recommendation system. The digital avatar recommendation system includes a computer-simulated digital avatar. An analog program of the digital avatar includes an agent implemented through reinforcement learning. In context, this is sometimes simply stated as: a digital avatar has an agent (obtained through reinforcement learning). It should be understood that the digital avatar recommendation system can be implemented by any apparatus, device, platform, device cluster having a computing and processing capabilities.

[0047] As shown in FIG. 3, the recommendation method includes the following steps: Step 31: The recommendation system obtains current state data, where the state data includes user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar. Step 32: The agent in the digital avatar maps the state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, where a candidate action in the candidate action set corresponds to a to-be-recommended content category, and the target action corresponds to a target content category. Step 33: The digital avatar performs target interaction with the target user, where the target interaction is used to recommend the target content category. The following describes specific execution of the above steps.

[0048] First, in step 31, the recommendation system obtains the current state data for the target user, that is, a current state s of an execution environment. The state data s can include the scenario information of the current scenario, e.g., in which the target user is located, the user information

of the target user, and a history of an interaction between the target user and the digital avatar.

[0049] In an implementation, the digital avatar and the target user are located in a scenario in a metaverse, for example, a store, a cinema, or a social occasion, where the digital avatar is a system role in the metaverse, and the target user can be a real human user or a player role controlled by the real human user. In this case, the scenario information can include information about the scenario in the metaverse, for example, a store location and a category, or a cinema location and film programming. In addition, the scenario information can further include information about an association relationship between the scenario and another scenario in the metaverse, etc. There are possibly a plurality of digital avatars in the metaverse scenario. In this case, the digital avatar described in the present specification can be any digital avatar suitable for recommendation in the metaverse.

[0050] In an implementation, the current scenario is a live streaming scenario, where the digital avatar is a live streamer in the live streaming scenario, and the target user is a user entering the live streaming scenario. In this case, the scenario information can include information related to the live streaming, such as an ID of the live streaming, a theme of the live streaming, and a layout of the live streaming. In a case of a live streaming dedicated to introduce/purchase commodities opened by a store, e.g., , a live streaming for sending commodities, information about the live streaming can include category information of the store to which the live streaming belongs, a commodity layout displayed in the live streaming, an element in the live streaming with which the target user has interacted (for example, clicked or browsed), etc.

[0051] The digital avatars and the target users in the above various scenarios can be in a one-to-one relationship, or can be in a one-to-many relationship. In a one-to-one scenario, one digital avatar interacts with only one user at a same time. In a one-to-many scenario, one digital avatar can interact with a plurality of users at a same time. For example, in a typical live streaming scenario, a digital avatar in live streaming can interact with a plurality of users entering the live streaming, and the target user can be any one of the plurality of users. However, in a case of targeted recommendation in the present specification, the plurality of users entering the live streaming can experience different recommendation interactions.

[0052] To provide individualized recommendation for the target user, the state data s further includes the user information of the target user, which reflects a feature of the target user. The user information can include basic attribute features of the target user, such as a user identifier, a

nickname, a user profile picture, registration duration, an education level, and an occupation, and can further include a user profile feature such as a group label to which the target user belongs. In addition, the user information can further include information about the rights and interests of the target user, for example, categories of discount cards, discount rates, and coupons that are owned
5 by the target user.

[0053] In an implementation, the user information can further include a user relationship network graph, or relationship information that is related to the target user and that is obtained based on the user relationship network graph, e.g., some graph information in the user relationship network graph. The user relationship network graph can be in a form of a user social relationship
10 graph, a user transaction relationship graph, etc. In such a user relationship network graph, a node represents a user, and an edge represents an association relationship between users, such as a social relationship or a transaction relationship. A more comprehensive embedding representation of the target user can be obtained based on adjacency information of the target user by using the user relationship network graph.

[0054] In addition to the static user information, the state data s further includes the history information of the interaction between the target user and the digital avatar. In an example, the interaction history can include a conversation record between the target user and the digital avatar, and the conversation record can be in a text form, a voice form, etc. The interaction history can further include historical recommendation information of the target user, for example, a
20 recommended content list of content that the digital avatar has recommended to the target user, and an accepted content list of content that has been recommended by the digital avatar and accepted by the target user. Based on different recommended content, a behavior of accepting the recommended content can be: clicking the recommended content; browsing the recommended content for more than a threshold duration; when the recommended content is information about
25 rights and interests (for example, a red envelope or a coupon), obtaining the rights and interests in the recommended content; or when the recommended content is a commodity, buying the commodity, etc. In some implementations, the historical recommendation information can further include a rejected content list of content that the target user has explicitly rejected, a content of interest list of recommended content that the target user is interested in (for example, adding the
30 recommended content to favorites, or setting the recommended content to "see later"), etc.

[0055] To represent the current environment state more comprehensively, in an

implementation, knowledge graph information related to the to-be-recommended content is further included in the state data. For example, an entity in a knowledge graph can correspond to a to-be-recommended content category, and a relationship between entities corresponds to a relationship between to-be-recommended content. For example, the to-be-recommended content is clothing commodities. Entities can correspond to categories such as hats, scarves, upper clothes, trousers, dresses, shoes. A relationship between these entities can be a relationship between these categories, for example, a matching relationship, an associated purchase, or a shared discount.

[0056] According to an example implementation, the digital avatar has configuration parameters defining character features of the digital avatar. These configuration parameters can define individualized content such as a personality feature, an action feature, and a language style of the digital avatar. In some systems, all digital avatars share the configuration parameters. In some other systems, different digital people have different configuration parameters. In an implementation, the configuration parameters of the current digital avatar are also included in the state data s to better describe a current environment state of the recommendation system.

[0057] It should be appreciated that the state data includes a plurality of aspects of environment state information. As shown in FIG. 2, the plurality of aspects of information are often located in different modes or different forms. Therefore, the state data can be obtained by fusing source data of a plurality of modes, and the source data of the plurality of modes can include text source data, image source data, audio source data, video source data, etc. For example, the text source data can include a chat text record of the target user, the image source data can include a profile picture of the target user and a layout picture in a live streaming, the audio source data can include a voice of the target user, or the video source data can include a digital recorded screen of a process of interaction between the target user and the digital avatar. In some implementations, the state data can further include graph data. The graph data includes graph data of the user relationship network graph and/or graph data of the knowledge graph related to the recommended content. Through the above multimodal fusion, a more rich and comprehensive state representation can be obtained.

[0058] It should be noted that although instances of information items included in the state data are described above by using examples, the state data does not necessarily include content in all the above examples. In some implementations, the information items enumerated above can be selected or amplified based on a system requirement, so as to flexibly form the state data.

[0059] Next, in step 32, the current state data s is input to the agent in the digital avatar. The

agent maps the state data to the target action a in the candidate action set based on the current policy π obtained through reinforcement learning, e.g., determines a recommendation action a for the target user from the candidate action set A . It should be understood that the candidate action set A is used to define action space formed by all possible recommendation actions. In an implementation, candidate actions in the candidate action set A corresponds to to-be-recommended content categories, e.g., the content categories form candidate action space.

[0060] In an implementation, a content category corresponds to a to-be-recommended content item, such as a commodity, an article, a video, etc. In this case, a dimension of the action space is equal to a number of content items. For example, if a number of different commodities to be recommended/sold in a live streaming of a store is N , corresponding N candidate actions form the action space.

[0061] To reduce complexity, in an implementation, a content category corresponds to a type of content item. For example, to-be-recommended content items are classified in advance, and a content category is formed based on a type of content item obtained through classification. In this case, the dimension of the action space is equal to a number of content item categories. For example, if clothing commodities are recommended/sold in a live streaming of a store, and are divided into six categories in advance: hats, scarves, upper clothes, trousers, dresses, and shoes, corresponding six candidate actions form the action space.

[0062] It should be understood that a division granularity of the action space can be set based on factors such as differentiation of to-be-recommended content, required precision, and a computing capability.

[0063] As described herein, the policy π obtained through reinforcement learning can be considered as a function for mapping from environment state space to the action space. Therefore, the agent in the digital avatar can map the current state data s to the target action a in the candidate action set based on the current policy π . In an implementation, the agent can be implemented by a neural network.

[0064] According to an implementation, the agent can determine the target action as follows. First, long-term cumulative reward Q values corresponding to candidate actions in the candidate action set are determined based the state data s . As defined in the above equation (2), a Q value corresponding to an action i represents a cumulative reward in a t^{th} step brought by using the policy π after the action i is performed starting from a current state s . Then, the target action a is

determined based on the Q values corresponding to the candidate actions.

[0065] In an example, an action corresponding to a maximum value in the Q values can be determined as the target action based on a greedy algorithm. In an example, the target action is selected from actions corresponding to non-maximum values by using a probability ε , and an action corresponding to a maximum Q value is selected as the target action by using a probability $(1-\varepsilon)$.

[0066] It can be understood that the determined target action corresponds to the target content category. Based on the division granularity of the action space, the target content category can be a target content item or a type of content item.

[0067] In step 33, the digital avatar performs target interaction with the target user, and the target interaction is used to recommend the target content category.

[0068] For example, the digital avatar can perform the target interaction with the target user based on a feature of a scenario in which the digital avatar is located, so as to recommend the target content category corresponding to the target action. For example, a manner of the target interaction can be: sending recommended text to the target user, making a conversation with the target user in a text form or a voice form, displaying the target content category to the target user, presenting the target content category at a site in a live streaming, and introducing the target content category.

[0069] In an example, the digital avatar has configuration parameters reflecting character features of the digital avatar. In this case, the digital avatar can interact with the target user in a manner specific to the digital avatar based on the configuration parameters. For example, the digital avatar makes a voice conversation by using a specific tone and a voice intonation, adds an expression symbol commonly used by the digital avatar to conversation text, and displays the target content category by using a specific body action.

[0070] As such, the digital avatar can provide individualized and targeted recommendation for the target user in an interactive manner based on a feature of the target user.

[0071] It can be understood that when the target action is performed on a system environment, an environment state is changed. For example, a recommended list of a user is updated to be used for next round of interactive recommendation.

[0072] Further, depending on a feedback of the target user for the target action, for example, accepting the recommendation, refusing the recommendation, partially accepting the recommendation, adding the recommendation to favorites, the target action can correspond to an

instant reward score r . Therefore, in an implementation, after the digital avatar performs a target recommendation action for the target user, the method further includes step 34 of updating the policy based on the reward score. For example, in this step, the system obtains a target feedback of the target user for the target interaction; determines, based on the target feedback, a reward score
5 corresponding to the target action; and updates the current policy based on the reward score.

[0073] In an implementation, it is determined that the reward score is a first value if the target feedback indicates that the target user accepts the target content category; or it is determined that the reward score is a second value if the target feedback indicates that the target user does not accept the target content category, where the second value is less than the first value. For example,
10 if the target user accepts a current recommendation, for example, clicks/purchases a recommended commodity and watches a recommended video, the reward score is set to 1 (the first value); or if the user does not accept the current recommendation, the reward score is set to 0 (the second value). In a case of partial acceptance (for example, a plurality of commodities of one category are recommended at a time, and some of these commodities are accepted by the user, or the user adds
15 the recommended content to a shopping cart/adds the recommended content to favorites, but does not purchase/read the recommended content), more reward values can be further set based on an acceptance degree. Certainly, a higher acceptance degree of the target user for a recommendation action indicates a larger reward score.

[0074] As such, the obtained reward score can be used to update a policy. In an implementation,
20 updating the policy can be updating the current policy in the agent based on the currently obtained reward score r . It can be understood that, although the policy can also be updated once based on a feedback of a single recommendation action, e.g., a policy parameter is updated once for a plurality of feedback actions. For example, the current state data s , the target action a , the obtained reward score r , and a new state s' migrated to can form a piece of training data (s, a, r, s') , and the training
25 data is added to a training data set. When an amount of training data is accumulated, the policy π is updated once based on the training data that includes the current reward score r . The policy can be updated by using various reinforcement learning algorithms, which are not limited herein.

[0075] In the above process, the agent obtained through reinforcement learning is embedded into the digital avatar, and the agent determines the recommendation action based on the feature
30 of the target user, so that the digital avatar provides targeted and individualized recommendation for the target user in an interactive manner, to form a new digital avatar recommendation system.

[0076] According to an implementation of an aspect, a digital avatar recommendation system is provided. The system can be deployed in any device, platform, or a device cluster having data storage, computing, and processing capabilities. FIG. 4 is a schematic diagram illustrating a structure of a digital avatar recommendation system according to an implementation. As shown in
5 FIG. 4, the recommendation system 400 includes a state acquisition module 41 and a computer-simulated digital avatar 42. An agent 43 is embedded into the digital avatar.

[0077] The state acquisition module 41 is configured to obtain current state data, where the state data includes user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar.

10 **[0078]** The agent 43 is configured to map the state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, where a candidate action in the candidate action set corresponds to a to-be-recommended content category, and the target action corresponds to a target content category.

[0079] The digital avatar 42 is configured to perform target interaction with the target user,
15 where the target interaction is used to recommend the target content category.

[0080] According to an implementation, the digital avatar recommendation system 400 further includes an updating module 44. The updating module is configured to: obtain a target feedback of the target user for the target interaction; determine, based on the target feedback, a reward score corresponding to the target action; and update the current policy based on the reward score.

20 **[0081]** For implementations of the modules and components in the recommendation system, references can be made to the above descriptions provided with reference to FIG. 3. Through the above recommendation system, targeted and individualized recommendation can be better provided for a user in an interactive manner.

[0082] According to an implementation of another aspect, a computer-readable storage
25 medium is further provided. The computer-readable storage medium stores a computer program. When the computer program is executed in a computer, the computer is enabled to perform the method described with reference to FIG. 2 or FIG. 4.

[0083] According to an implementation of still another aspect, a computing device is further provided and includes a memory and a processor. The memory stores executable code, and when
30 executing the executable code, the processor implements the method described with reference to FIG. 2 or FIG. 4.

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

[0085] The technical characteristics, technical solutions, and beneficial effects of the present specification are further described in detail in the above illustrative implementations. It should be understood that the above descriptions are merely example implementations of the present specification, but are not intended to limit the protection scope of the present specification. Any
10 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 recommendation method, performed by a digital avatar recommendation system, the digital avatar recommendation system comprising a computer-simulated digital avatar, and the method comprising:

obtaining current state data, the state data including user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar;

mapping, by an agent in the digital avatar, the current state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, a candidate action in the candidate action set corresponding to a to-be-recommended content category, and the target action corresponding to a target content category; and

performing, by the digital avatar, a target interaction with the target user, wherein the target interaction is used to recommend the target content category.

2. The method according to claim 1, wherein the current scenario is a live streaming scenario, the digital avatar is a live streamer in the live streaming scenario, and the target user is a user entering the live streaming scenario.

3. The method according to claim 2, wherein the scenario information includes a theme of the live streaming, category information of a store to which the live streaming belongs, and an element in the live streaming with which the target user has interacted.

4. The method according to claim 1, wherein the current scenario is a metaverse, and the digital avatar is a system role in the metaverse.

5. The method according to claim 1, further comprising:

obtaining a target feedback of the target user for the target interaction;

determining, based on the target feedback, a reward score corresponding to the target action;

and

updating the current policy based on the reward score.

6. The method according to claim 5, wherein the determining the reward score corresponding to the target action including:

5 determining that the reward score is a first value if the target feedback indicates that the target user accepts the target content category; or

 determining that the reward score is a second value if the target feedback indicates that the target user does not accept the target content category, where the second value is less than the first value.

10

7. The method according to claim 1, wherein the mapping the state data to the target action in the candidate action set includes:

 determining, based on the state data, long-term cumulative reward Q values corresponding to candidate actions in the candidate action set, to obtain Q values; and

15 determining the target action based on the Q values.

8. The method according to claim 1, wherein the current state data is obtained by fusing source data of a plurality of modes, and the source data of the plurality of modes includes at least two of: text source data, image source data, audio source data, or video source data.

20

9. The method according to claim 1, wherein the user information includes relationship information obtained based on a user relationship network graph, and the current state data includes graph data of the relationship network graph.

25 10. The method according to claim 1, wherein the current state data further includes graph data of a knowledge graph, and an entity in the knowledge graph corresponds to the to-be-recommended content category.

30 11. The method according to claim 1, wherein the digital avatar has configuration parameters defining character features of the digital avatar, and the state data further includes the configuration parameters.

12. A digital avatar recommendation system, comprising a state acquisition module and a computer-simulated digital avatar, an agent being embedded into the digital avatar;

the state acquisition module being configured to obtain current state data, the state data including user information of a target user, scenario information of a current scenario, and history information of an interaction between the target user and the digital avatar;

the agent being configured to map the state data to a target action in a candidate action set based on a current policy obtained through reinforcement learning, a candidate action in the candidate action set corresponding to a to-be-recommended content category, and the target action corresponding to a target content category; and

the digital avatar being configured to perform target interaction with the target user, wherein the target interaction is used to recommend the target content category.

13. The digital avatar recommendation system according to claim 12, further including an updating module, wherein the updating module is configured to obtain a target feedback of the target user for the target interaction;

determine, based on the target feedback, a reward score corresponding to the target action; and

update the current policy based on the reward score.

14. 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 11.