RECOMMENDATION METHOD AND APPARATUS FOR MEDIA OBJECT, DEVICE AND MEDIUM

RELATED APPLICATION

[0001] This application claims priority to Chinese Patent Application No. 2024117889825, entitled "DATA PROCESSING METHOD AND APPARATUS FOR MEDIA OBJECT, DEVICE AND MEDIUM" filed with the China National Intellectual Property Administration on December 5, 2024, which is incorporated herein by reference in its entirety.

FIELD OF THE TECHNOLOGY

[0002] The present disclosure relates to the field of Internet technologies, and in particular, to a recommendation method and apparatus for a media object, a device, and a medium.

BACKGROUND OF THE DISCLOSURE

[0003] Currently, cold-start recommendation is a case of recommending a currently obtained cold-start media object (for example, a newly introduced media object, where the newly introduced media object may be specifically a currently collected media object with no interaction data amount and/or with only a small interaction data amount) in a recommendation system. The newly introduced media object (for example, a livestreaming room created by a streamer user) lacks corresponding historical interaction behavior data. Therefore, when performing cold-start recommendation on the media object through a recommendation model, the recommendation system usually recommends a similar label based on an object label of the media object. For example, for a livestreaming room 1 with a media object of a category-A style, when livestreaming recommendation needs to be performed through the recommendation model, the recommendation system directly searches, based on a style label (namely, the category-A style) of the livestreaming room 1, massive livestreaming rooms for another livestreaming room (for example, a livestreaming room 2) that carries the style label (namely, the category-A style), and then may push the livestreaming room 1 to a service object (namely, a viewer user) that is currently watching a live video in the livestreaming room 2, to implement cold-start recommendation for the livestreaming room 1.

[0004] However, when cold-start recommendation is performed based on a label of a media object, because a live video (for example, a live video 1) in a livestreaming room 1 lacks actual interaction record information, the livestreaming room 1 can be directly pushed, based on viewer

users recorded in interaction record information of a live video (for example, a live video 2) in a livestreaming room 2, to the viewer objects when the livestreaming room 2 is found through matching between similar labels. However, when a style label of the live video 2 is different from actual live content in the live video 2, live content in the pushed livestreaming room 1 can hardly truly meet viewing requirements of the viewer users. Consequently, accuracy of performing media recommendation for some media objects with a small interaction data amount is reduced to some extent.

SUMMARY

[0005] Embodiments of the present disclosure provide a recommendation method and apparatus for a media object, a device, and a medium, to improve media recommendation accuracy for a cold-start media object (to be specific, a media object with a small interaction data amount).

[0006] An aspect of the embodiments of the present disclosure provides a recommendation method for a media object, including:

[0007] obtaining target audio data of a target media object to be recommended in a target sampling window, and performing text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the target media object;

[0008] obtaining N historical text representation vectors for representing N historical media objects, and performing vector similarity calculation on the target text representation vector and each of the N historical text representation vectors respectively, to obtain N vector similarities, an interaction data amount of each of the N historical media objects being greater than an interaction data amount of the target media object, N being an integer greater than 1, each of the N historical text representation vectors being recorded after text vector representation processing is performed on historical speech recognition text corresponding to each piece of historical audio data, each piece of historical audio data being audio data of each historical media object in a recorded historical sampling window, and window duration of the historical sampling window being consistent with window duration of the target sampling window;

[0009] selecting, based on the N vector similarities, K historical media objects from the N historical media objects, and establishing a similarity association relationship between the K historical media objects and the target media object, K being a positive integer less than N; and

[0010] obtaining K historical recommendation features of the K historical media objects, and determining, based on the similarity association relationship and K historical recommendation

features when the K historical recommendation features of the K historical media objects are obtained, a target recommendation feature for representing the target media object, and recommending the target media object based on the target recommendation feature.

[0011] An aspect of the embodiments of the present disclosure provides a recommendation apparatus for a media object, including:

[0012] an audio data speech recognition module, configured to obtain target audio data of a target media object to be recommended in a target sampling window, and perform text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the target media object;

[0013] a vector similarity calculation module, configured to obtain N historical text representation vectors for representing N historical media objects, and perform vector similarity calculation on the target text representation vector and each of the N historical text representation vectors respectively, to obtain N vector similarities, an interaction data amount of each of the N historical media objects being greater than an interaction data amount of the target media object, N being an integer greater than 1, each of the N historical text representation vectors being recorded after text vector representation processing is performed on a historical speech recognition text corresponding to each piece of historical audio data, each piece of historical audio data being audio data of each historical media object in a recorded historical sampling window, and a window duration of the historical sampling window being consistent with a window duration of the target sampling window;

[0014] a similarity association relationship establishment module, configured to select, based on the N vector similarities, K historical media objects from the N historical media objects, and establish a similarity association relationship between K historical media objects and the target media object, K being a positive integer less than N; and

[0015] a target recommendation feature determining module, configured to obtain K historical recommendation features of the K historical media objects, and determine, based on the similarity association relationship and the K historical recommendation features, a target recommendation feature for representing the target media object, and recommend the target media object based on the target recommendation feature.

[0016] An aspect of the embodiments of the present disclosure provides a computer device, including a memory and a processor. The memory is connected to the processor. The memory is configured to store a computer program. The processor is configured to invoke the computer

program, to enable the computer device to perform the method provided in the foregoing aspect of the embodiments of the present disclosure.

[0017] An aspect of the embodiments of the present disclosure provides a computer-readable storage medium. The computer-readable storage medium has a computer program stored therein. The computer program being adapted to be loaded and executed by a processor, to enable a computer device with the processor to perform the method provided in the foregoing aspect of the embodiments of the present disclosure.

[0018] An aspect of the embodiments of the present disclosure provides a computer program product or a computer program. The computer program product or the computer program includes computer instructions, and the computer instructions are stored in a computer-readable storage medium. A processor of a computer device reads the computer instructions from the computer-readable storage medium, and the processor executes the computer instructions, to enable the computer device to perform the method provided in the foregoing aspect.

[0019] When obtaining target audio data of a target media object to be recommended (for example, a current livestreaming room) in a target sampling window, the computer device in the embodiments of the present disclosure may perform speech recognition on the target audio data in a speech recognition mode to obtain target speech recognition text corresponding to the target audio data. The target media object herein is a media object with a smaller interaction data amount relative to interaction data amounts of N historical media objects (for example, N historical livestreaming rooms with large interaction data amounts). To be specific, an interaction data amount of each of the N historical media objects is greater than the interaction data amount of the target media object. Therefore, in the embodiments of the present disclosure, when a target text representation vector for representing the target media object is obtained by performing text vector representation processing on the target speech recognition text in a text vector representation processing mode, N historical text representation vectors for representing the N historical media objects may be further obtained, and then vector similarity calculation is separately performed on the target text representation vector and each of the N historical text representation vectors to obtain N vector similarities. A historical text representation vector of any one of the N historical media objects is recorded after text vector representation processing is performed on historical speech recognition text of a corresponding historical media object. To be specific, one historical text representation vector is recorded after text vector representation processing is performed on historical speech recognition text corresponding to one piece of historical audio data, and one piece of historical audio data is audio data of one historical media object in a recorded historical sampling

window. Window duration of the historical sampling window herein is consistent with window duration of the target sampling window. Further, the computer device in the embodiments of the present disclosure may select K historical media objects from the N historical media objects through vector similarity search based on the N vector similarities (to be specific, all of the selected K historical media objects are media objects with high vector similarities to the target media object, for example, K historical livestreaming rooms with large interaction data amounts). In this way, in the embodiments of the present disclosure, the K historical media objects with high vector similarities may be configured for representing the target media object, to directly (namely, explicitly) establish a similarity association relationship between the target media object and the K historical media objects. Based on this, in the embodiments of the present disclosure, when K historical recommendation features of the K historical media objects are obtained, a target recommendation feature of the target media object (for example, the current livestreaming room) may be further accurately determined based on the similarity association relationship. In this way, when media recommendation is performed for the target media object based on the target recommendation feature, accuracy of the media recommendation can be improved, and a technical problem in the related art that accuracy of media recommendation is affected by a lack of actual interaction record information of the target media object (namely, a cold-start media object, for example, a cold-start livestreaming room) with a small interaction data amount can also be resolved from the source. In other words, in the embodiments of the present disclosure, when the target media object is cold-start target media, the K historical media objects (for example, the K historical livestreaming rooms with large interaction data amounts) that have high vector similarities and that are determined through vector similarity search may be further configured for explicitly representing the cold-start media object, to boost recommendation of the cold-start media object.

BRIEF DESCRIPTION OF THE DRAWINGS

[0020] FIG. 1 is a diagram of a structure of a network architecture according to an embodiment of the present disclosure.

[0021] FIG. 2 is a diagram of a data exchange scenario according to an embodiment of the present disclosure.

[0022] FIG. 3 is a schematic flowchart of a recommendation method for a media object according to an embodiment of the present disclosure.

[0023] FIG. 4 is a diagram of an interaction data curve of historical media objects according to an embodiment of the present disclosure.

[0024] FIG. 5 is a diagram of a similarity association relationship according to an embodiment of the present disclosure.

[0025] FIG. 6 is a diagram of a scenario of recommending a livestreaming media object on a service processing platform according to an embodiment of the present disclosure.

[0026] FIG. 7 is a diagram of a scenario of livestreaming in a livestreaming room according to an embodiment of the present disclosure.

[0027] FIG. 8 is a diagram of a scenario of recommending a video-on-demand media object on a service processing platform according to an embodiment of the present disclosure.

[0028] FIG. 9 is a diagram of a scenario of playing a video based on a video-on-demand identifier according to an embodiment of the present disclosure.

[0029] FIG. 10 is a diagram of a recommendation display list of a service processing platform according to an embodiment of the present disclosure.

[0030] FIG. 11 is a flowchart of a processing method for invalid target audio according to an embodiment of the present disclosure.

[0031] FIG. 12 is a flowchart of another processing method for invalid target audio according to an embodiment of the present disclosure.

[0032] FIG. 13 is a schematic flowchart of estimating a target service object for a media object according to an embodiment of the present disclosure.

[0033] FIG. 14 is a schematic flowchart of estimating a target service object for a livestreaming room according to an embodiment of the present disclosure.

[0034] FIG. 15 is a schematic flowchart of another recommendation method for a media object according to an embodiment of the present disclosure.

[0035] FIG. 16 is a diagram of dynamic sampling based on a play timeline of to-be-processed livestreaming data according to an embodiment of the present disclosure.

[0036] FIG. 17 is a diagram of a structure of a recommendation apparatus for a media object according to an embodiment of the present disclosure.

[0037] FIG. 18 is a diagram of a structure of a computer device according to an embodiment of the present disclosure.

DESCRIPTION OF EMBODIMENTS

[0038] Concepts of related technical terms in the embodiments of the present disclosure are

described below:

[0039] 1. Media object: A media object in the embodiments of the present disclosure includes a to-be-processed media object (namely, a to-be-recommended media object) and a recommended historical media object. The to-be-processed media object (namely, the to-be-recommended media object) and the historical media object herein are media objects of a same type. The media object herein is a first media object (for example, a virtual livestreaming room, a conference room, or a classroom) with real-time video processing or a second media object (for example, an on-demand video) with flexible video progress adjustment that is displayed on an application client. A video in the first media object may include, but is not limited to, a live video, a conference video, and an educational video. When the media object is the to-be-recommended media object, the first media object herein may specifically include a to-be-recommended livestreaming media object in a real-time livestreaming scenario. Similarly, the second media object may specifically include a to-be-recommended video-on-demand media object that is obtained by performing live recording in the real-time livestreaming scenario and that currently supports video replay.

[0040] The live video may specifically include a real-time live video displayed in a livestreaming room (namely, the first media object) on a livestreaming client in the real-time livestreaming scenario and/or a recorded video displayed in a livestreaming room (namely, the second media object) on the livestreaming client in a live replay scenario. The on-demand video may specifically include a long video and a short video that are displayed on a video play client in a video-on-demand scenario. The long video is a video with video duration greater than a video duration threshold, and the short video is a video with video duration less than the video duration threshold.

[0041] In the embodiments of the present disclosure, media objects that have small interaction data amounts and that need boosted recommendation in a recommendation system may be collectively referred to as the to-be-processed media object or the to-be-recommended media object (for example, the to-be-recommended media object may be some media objects that are statistically collected by the recommendation system and that have small volumes of interaction data with a user), and media objects that have large interaction data amounts and for which media recommendation has been performed in the recommendation system are collectively referred to as the historical media object (for example, the historical media object may be some media objects that are statistically collected by the recommendation system and that have large volumes of interaction data with the user). Specifically, for example, in the real-time livestreaming scenario, a cold-start

livestreaming room that currently needs boosted recommendation or a tail livestreaming room satisfying long-tail distribution (that is, a long-tail livestreaming room) may be referred to as a to-be-recommended media object, and some popular livestreaming rooms for which livestreaming or livestreaming recommendation is completed or head livestreaming rooms satisfying long-tail distribution may be referred to as historical media objects. In other words, in the real-time livestreaming scenario, the to-be-recommended media object in the embodiments of the present disclosure may specifically include a cold-start livestreaming room (for example, a newly introduced livestreaming room) and a tail livestreaming room satisfying long-tail distribution (that is, a long-tail livestreaming room, for example, some livestreaming rooms that are not newly introduced but have small amounts of interaction data), and the historical media object in the embodiments of the present disclosure may specifically include some popular livestreaming rooms with large interaction data amounts and head livestreaming rooms satisfying long-tail distribution. The interaction data amount (namely, a click-through volume) herein is a data volume, cumulatively collected by the recommendation system, of interaction operations such as clicks/taps and/or views performed by one or more users on a corresponding media object (or multimedia data of a corresponding media object). In the embodiments of the present disclosure, a size of the interaction data amount is determined relative to a preset interaction data amount threshold configured in an inference system. For example, a computer device in the recommendation system determines a livestreaming room with a currently collected interaction data amount less than the interaction data amount threshold as a livestreaming room with a small interaction data amount (that is, a low-interaction livestreaming room), and determines a livestreaming room with a currently collected interaction data amount greater than or equal to the interaction data amount threshold as a livestreaming room with a large interaction data amount (that is, a high-interaction livestreaming room).

[0042] In the embodiments of the present disclosure, a larger interaction data amount indicates higher popularity of a media object corresponding to the interaction data amount. On the contrary, a smaller interaction data amount indicates lower popularity of a media object corresponding to the interaction data amount. For example, for media objects such as a cold-start livestreaming room or a long-tail livestreaming room, interaction data amounts of the media objects that are cumulatively collected by the recommendation system are usually less than the interaction data amount threshold. The media objects have small interaction data amounts, and therefore lack historical behavior interaction data information for modeling an item relative to a user preference. Therefore, if attributive features of items of the media objects are directly used, distribution of the items in the low-popularity media objects may be delayed in the recommendation system.

Therefore, in the embodiments of the present disclosure, to improve timeliness and accuracy of distribution of the items in the low-popularity media objects, when a plurality of other recommended historical media objects that have high semantic similarities and large interaction data amounts and that match the media object with a small interaction data amount (that is, the to-be-recommended media object) are found, a recommendation feature (that is, a target recommendation feature) of the media object with a small interaction data amount (that is, the to-be-recommended media object) can be equivalently represented by recommendation features (that is, historical recommendation features) of the plurality of other recommended historical media objects that have high semantic similarities and large interaction data amounts, to implement boosted recommendation for the media object with a small interaction data amount (that is, the to-be-recommended media object).

[0043] 2. Sampling window: A sampling window in the embodiments of the present disclosure includes a target sampling window and a historical sampling window. Both the target sampling window and the historical sampling window herein are windows for defining a video signal sampling parameter during video capturing. The video signal sampling parameter specifically includes a sampling frequency, a resolution, a frame rate, and the like.

[0044] In the embodiments of the present disclosure, in a service scenario (for example, a real-time livestreaming scenario), sampling windows of multimedia data (for example, to-be-recommended livestreaming data) of a to-be-recommended media object within specific duration (that is, preset window duration) that is captured backward in real time from a current sampling time point (for example, a time point currently indicating a latest livestreaming progress of a livestreaming room) based on a sampling frequency are collectively referred to as the target sampling window, and sampling windows of multimedia data of a historical media object within specific duration that is historically captured based on the same sampling frequency are collectively referred to as the historical sampling window. Specifically, in the real-time livestreaming scenario, a sampling window for dynamically capturing livestreaming data within specific duration with proceeding of a livestreaming progress of a livestreaming room may be referred to as the target sampling window. To be specific, in the embodiments of the present disclosure, when different sampling end time points are dynamically determined through dynamic sampling based on different livestreaming progresses, sampling windows corresponding to the different sampling end time points may be obtained. Therefore, a latest sampling window determined from the sampling windows corresponding to the different sampling end time points may be dynamically determined as the target sampling window. Then different audio data in the to-be-recommended media object may be obtained through dynamic sampling based on the target sampling window, so that redundant

noise data can be subsequently removed from currently sampled audio data through noise removal, to obtain target audio data related to a speaker. In addition, when livestreaming is completed, the last sampling window of all sampling windows in which livestreaming is currently completed and sampling is completed may be referred to as the historical sampling window.

[0045] The target sampling window herein is a latest sampling window selected from a series of sampling windows for a same media object (that is, a same to-be-recommended media object) that currently needs boosted recommendation. A video clip and/or an audio clip with its duration being the preset window duration may be captured, based on the latest sampling window, from a to-be-recommended media object (for example, a livestreaming room or an on-demand video) that has a large amount of media information, and then target audio data (for example, audio data related to speech of a speaker) in the target sampling window may be extracted from the captured video clip or audio clip.

[0046] 3. Audio data: Audio data in the embodiments of the present disclosure includes target audio data and historical audio data. Both the target audio data and the historical audio data are audio data (which may be, for example, audio encoded data represented in a binary form) obtained by encoding a corresponding analog sound signal into a digital form through analog-to-digital conversion. In the embodiments of the present disclosure, when multimedia data of a to-be-recommended media object is captured based on a target sampling window (for example, to-be-recommended livestreaming data such as a video clip or an audio clip is captured), audio data related to speech of a speaker and extracted from the multimedia data of the to-be-recommended media object may be referred to as the target audio data. In this case, the target audio data may be obtained by denoising audio data in the video clip or the audio clip in the target sampling window. Similarly, in the embodiments of the present disclosure, when multimedia data (for example, historical livestreaming data) of a historical media object is captured based on a historical sampling window, audio data extracted from the multimedia data of the historical media object may be referred to as the historical audio data.

[0047] In some embodiments, for example, in a real-time livestreaming scenario, audio data in a video clip or an audio clip captured from a current livestreaming room based on a target sampling window is referred to as the target audio data, and audio data in a video clip or an audio clip captured, based on a historical sampling window, from a historical livestreaming room in which livestreaming is currently completed may be referred to as the historical audio data.

[0048] 4. Similarity search: A similarity search in the embodiments of the present disclosure is a mode of performing vector similarity search based on a vector similarity between different

search objects. In the embodiments of the present disclosure, for a given search object, (that is, the foregoing to-be-recommended media object), a recommendation system may search, through vector similarity search, a preset search object set (that is, a set including N historical media objects) for other search objects (that is, the foregoing K historical media objects) similar to or matching the search object (that is, the foregoing to-be-recommended media object).

[0049] In the embodiments of the present disclosure, during a similarity search in the vector similarity search mode, similarity calculation needs to be first performed on a target text representation vector and a historical text representation vector of each historical media object in the search object set to obtain a vector similarity between the target text representation vector and each historical text representation vector, and then a historical media object with a high vector similarity to the to-be-recommended media object may be accurately found from the search object set based on the obtained vector similarity. Specifically, for example, in a real-time livestreaming scenario, vector similarity calculation may be performed on a target text representation vector of a current livestreaming room (namely, the to-be-recommended media object) and a historical text representation vector of each historical livestreaming room (namely, each historical media object) to obtain a vector similarity between the current livestreaming room (namely, the to-be-recommended media object) and each historical livestreaming room (namely, each historical media object). In this case, the recommendation system may further sort the historical media objects (for example, N historical livestreaming rooms) in descending order of the vector similarities obtained through calculation. Then a sorting list including the sorted historical media objects may be collectively referred to as a historical media object list. Therefore, top K (that is, K=20) historical media objects (for example, K historical livestreaming rooms) that are ranked top may be selected from the historical media object list as historical media objects associated with the to-be-recommended media object (for example, historical livestreaming rooms with high vector similarities to the current livestreaming room).

[0050] 5. Recommendation feature: A recommendation feature in the embodiments of the present disclosure includes a target recommendation feature and a historical recommendation feature. The recommendation feature herein is feature data for describing attribute data of a user and/or an item in a recommendation system. The feature data may help the recommendation system understand preferences of a service object (for example, a user U1) for some items (for example, live videos related to interpretation of some items of a specific theme in some livestreaming rooms). Therefore, when a related item is recommended to the service object (for example, the user U1), a live clip in a livestreaming room to which the item belongs may be recommended to the user U1, to enhance recommendation accuracy and help the user U1 find an item satisfying a preference of the

user U1.

[0051] In the embodiments of the present disclosure, feature data that can describe attributive data corresponding to a to-be-recommended media object is referred to as the target recommendation feature, and feature data that can describe attributive data corresponding to a historical media object is referred to as the historical recommendation feature. Specifically, for example, in a real-time livestreaming scenario, in the embodiments of the present disclosure, attribute data corresponding to historical media data may be extracted from historical behavior interaction information of the historical media object through a recommendation model, vectorized feature calculation is further performed on the attribute data, and then feature data obtained through feature calculation may be collectively referred to as the historical recommendation feature of the historical media object. In this way, when obtaining, through calculation, a historical recommendation feature of each of K historical media objects through the recommendation model, the recommendation system may further directly pool K historical recommendation features through pooling processing based on a similarity association relationship established between the K historical media objects and the to-be-recommended media object, to equivalently obtain, by using a pooling feature obtained through pooling processing, the feature data (that is, the target recommendation feature) for describing the attribute data corresponding to the to-be-recommended media object.

[0052] In the embodiments of the present disclosure, the pooling processing may include at least average pooling processing, max pooling processing, and random pooling processing. The pooling processing herein is a special operation in a convolutional neural network (CNN), and is mainly configured for extracting, from a specific region, key information of the region. For example, in the embodiments of the present disclosure, in a region including the K historical recommendation features, key information that can fully represent the region may be extracted as a pooling feature (namely, the pooling processing feature) through pooling processing. Then the target recommendation feature for representing the to-be-recommended media object may be equivalently obtained by using the fully extracted pooling feature (namely, the pooling processing feature), to subsequently implement media recommendation (for example, cold-start recommendation) for the to-be-processed media object (namely, the to-be-recommended media object) in the recommendation system by using the target recommendation feature in a collaborative filtering mode. The CNN herein may be a network unit in the recommendation model, or may be a network unit independent of the recommendation model. A specific deployment mode of the CNN for pooling processing is not limited in the embodiments of the present disclosure. In addition, a specific implementation of pooling processing is not limited in the embodiments of the present

disclosure either.

[0053] The average pooling processing is a processing mode in which a computer device may perform an averaging operation on the region including the K historical recommendation features. In this way, the computer device may further use an average value (that is, a pooling processing feature) obtained through the averaging operation as the target recommendation feature. Similarly, the max pooling processing is another processing mode in which the computer device may retain a maximum historical recommendation feature (that is, a historical recommendation feature with a largest feature value) in the region including the K historical recommendation features. In this way, the computer device may further use the maximum historical recommendation feature (that is, another pooling processing feature) retained in the region as the target recommendation feature. By analogy, the random pooling processing is still another processing mode in which the computer device may perform random selection from the K historical recommendation features in the region including the K historical recommendation features based on a random selection probability separately configured for each historical recommendation feature. For example, in the region including the K historical recommendation features, a historical recommendation feature with a larger feature value is configured with a higher random selection probability; and on the contrary, a historical recommendation feature with a smaller feature value is configured with a lower random selection probability. In this way, in the embodiments of the present disclosure, after weighted calculation is separately performed on a corresponding historical recommendation feature in a weighted calculation mode based on the random selection probability configured for each historical recommendation feature, random selection may be further performed from historical recommendation features with corresponding random selection probabilities for weighted calculation. In this way, the computer device may further use, as the target recommendation feature, a historical recommendation feature (that is, still another pooling processing feature) that is randomly selected from the region and that has a corresponding random selection probability for weighting calculation.

[0054] 6. Service processing platform: A service processing platform in the embodiments of the present disclosure is a software platform on which a target recommendation model is deployed. The service processing platform is a digital platform supporting service operations and management activities of an enterprise or an organization.

[0055] In the embodiments of the present disclosure, software platforms provided by a recommendation system to perform media recommendation and media display for a media object (that is, the foregoing to-be-recommended media object) that needs boosted recommendation may

be collectively referred to as the service processing platform. Specifically, for example, in a real-time livestreaming scenario, when the recommendation system needs to perform boosted recommendation for livestreaming data of a livestreaming room, the recommendation system may add and display, on the service processing platform based on a matching essence between a user and an item, the livestreaming room that currently needs boosted recommendation. In this way, the recommendation system may determine, based on the matching essence between a user and an item, a user matching the to-be-recommended media object (namely, the livestreaming room), and then the user matching the to-be-recommended object may be collectively referred to as a target service object. Based on this, when the target service object accesses the service processing platform, the recommendation system may push a recommendation display list including the to-be-recommended media object to the target service object, and then may display (for example, play) the livestreaming data of the livestreaming room to the target service object in the recommendation display list. The service processing platform in the embodiments of the present disclosure includes, but is not limited to, a social software platform.

[0056] 7. Service object: A service object in the embodiments of the present disclosure includes all service objects accessing the service processing platform. The service objects include a target user (namely, a target service object) that matches the to-be-recommended media object and that is determined by a recommendation system from all of the service objects based on a matching essence between a user and an item (that is, user-item). The service object is a software abstraction of an entity in the real world, and may represent a person in a specific service field. For example, in the social field, that is, the social software platform, the service object in the embodiments of the present disclosure may represent any user accessing the service processing platform, or may represent a target user (namely, a target service object) that is determined by a recommendation model through analytical processing and that is configured for receiving a media object recommended by the recommendation system.

[0057] In the embodiments of the present disclosure, all users having access to the service processing platform are referred to as service objects, and online users matching the to-be-recommended media object are referred to as target service objects. Specifically, for example, in a real-time livestreaming scenario, in the embodiments of the present disclosure, all users on the service processing platform to which the livestreaming data of the livestreaming room that needs boosted recommendation belongs are referred to as service objects, and an online user matching the livestreaming room (which may be specifically the livestreaming data of the livestreaming room) is referred to as a target service object. In this way, the recommendation system may invoke a target recommendation model to recommend the livestreaming room to the target service object, to

implement boosted recommendation for the livestreaming data of the livestreaming room.

[0058] 8. Boosted recommendation: Boosted recommendation in the embodiments of the present disclosure is a recommendation means of equivalently representing a recommendation feature (that is, a target recommendation feature) of a media object with a small interaction data amount (that is, a to-be-recommended media object) by using a recommendation feature (that is, a historical recommendation feature) of another recommended historical media object that has a high semantic similarity and a large interaction data amount when the another recommended historical media object that has a high semantic similarity to the media object with a small interaction data amount (that is, the to-be-recommended media object) and that has a large interaction data amount is found. In other words, the boosted recommendation is to accurately recommend, to a target user by using a specific algorithm and technical means in the recommendation method of the present disclosure, content of a media object (for example, a cold-start livestreaming room or a long-tail livestreaming room) that lacks an interaction data amount and/or that has only a small interaction data amount, to increase an exposure rate and an interaction rate of the content.

[0059] The content of the cold-start livestreaming room or the long-tail livestreaming room may specifically include a video clip or an audio clip that is dynamically captured in a real-time video and/or audio recommendation scenario based on a target sampling window and in which a streamer interprets corresponding live content (for example, a livestreaming theme related to one or more items for which a user may trigger a transaction) in a livestreaming room. This means that, for livestreaming themes of a same media object in different sampling windows, the boosted recommendation method provided in the embodiments of the present disclosure can enable different livestreaming themes to dynamically match different target users, so that different livestreaming themes in different sampling windows can be pushed to different target users as much as possible, to increase exposure rates and interaction rates of video clips or audio clips related to different livestreaming themes.

[0060] In the embodiments of the present disclosure, during boosted recommendation, a to-be-recommended media object having a small volume of interaction data with a user may be recommended to a target user, to increase an exposure rate and an interaction rate of the to-be-recommended media object. Specifically, for example, in a real-time livestreaming scenario, a target recommendation model intelligently finds, through matching based on a target recommendation feature of the to-be-recommended media object (for example, the livestreaming room) that is obtained through equivalent calculation, a target service object (for example, a target user) matching the livestreaming room (which may be specifically, for example, a live video for interpreting an

item in the livestreaming room), to recommend the livestreaming room to the target user, so that the livestreaming room with boosted recommendation can be displayed in a recommendation display list of the target service object, to implement boosted recommendation for the livestreaming room.

[0061] FIG. 1 is a diagram of a structure of a network architecture according to an embodiment of the present disclosure. As shown in FIG. 1, the network architecture may include a server 10d and a user terminal cluster. The user terminal cluster may include one or more user terminals, and a quantity of user terminals is not limited herein. As shown in FIG. 1, the user terminal cluster may specifically include a user terminal 10a, a user terminal 10b, a user terminal 10c, a user terminal 20a, a user terminal 20b, a user terminal 20c, and the like. The server 10d may be an independent physical server, or may be a server cluster or a distributed system that includes a plurality of physical servers, or may be a cloud server that provides basic cloud computing services such as a cloud service, a cloud database, cloud computing, a cloud function, cloud storage, a network service, cloud communication, a middleware service, a domain name service, a security service, a content delivery network (CDN), big data, and an artificial intelligence platform. The user terminal 10a, the user terminal 10b, the user terminal 10c, the user terminal 20a, the user terminal 20b, the user terminal 20c, and the like each may include an electronic device with a video/image play function, for example, a smartphone, a tablet computer, a notebook computer, a palmtop computer, a mobile Internet device (MID), a wearable device (for example, a smartwatch or a smart band), an intelligent voice interaction device, an intelligent home appliance (for example, a smart television), or an in-vehicle device. As shown in FIG. 1, the user terminal 10a, the user terminal 10b, the user terminal 10c, the user terminal 20a, the user terminal 20b, the user terminal 20c, and the like each may be connected to the server 10d through a network connection, so that each user terminal can exchange data with the server 10d through the network connection.

[0062] For example, in a real-time livestreaming scenario, in the embodiments of the present disclosure, user terminals for capturing livestreaming data may be collectively referred to as a first local device. In this case, multimedia data (for example, to-be-recommended livestreaming data) of a to-be-recommended media object (for example, a to-be-recommended livestreaming media object) is captured by the first local device (for example, the user terminal 10a). Then the first local device (for example, the user terminal 10a) may transmit the captured multimedia data of the to-be-recommended media object to a cloud server. In this case, the cloud server (for example, the server 10d) may perform data processing on the received multimedia data of the to-be-recommended media object by using a recommendation method for a media object in the embodiments of the present disclosure, so that historical media objects matching the to-be-recommended media object (for example, the to-be-recommended livestreaming media object) can

be subsequently selected from a large quantity of historical media objects with large interaction data amounts through vector similarity search. Then a target recommendation feature of the to-be-recommended media object may be explicitly represented by historical recommendation features of the selected historical media objects. Then a target service object matching the multimedia data of the to-be-recommended media object may be finally found based on the target recommendation feature to perform media recommendation. To be specific, the cloud server (for example, the server 10d) may further push the to-be-recommended media object (which may be specifically the multimedia data of the to-be-recommended media object) to the target service object. For ease of distinguishing, in the embodiments of the present disclosure, a user terminal, in the user terminal cluster, that is used by the target service object is collectively referred to as a second local device. In this way, when the second local device receives the to-be-recommended media object (which may be specifically the multimedia data of the to-be-recommended media object) that is recommended by the cloud server (for example, the server 10d) through boosted recommendation, the target service object is enabled to view the multimedia data of the to-be-recommended media object on a service processing platform displayed on the second local device. In other words, the service processing platform (for example, a social software platform) in the embodiments of the present disclosure may be configured and run on another local device (that is, the second local device, for example, the user terminal 20a). In this way, when receiving the to-be-recommended media object (for example, the to-be-recommended livestreaming media object) that is recommended by the cloud server through boosted recommendation, the second local device (for example, the user terminal 20a) may display and play the multimedia data (for example, the to-be-recommended livestreaming data) of the to-be-recommended media object for the target service object on the service processing platform.

[0063] In the embodiments of the present disclosure, to accurately push the multimedia data of the to-be-recommended media object to the target service object to implement boosted recommendation for the to-be-recommended media object, the multimedia data of the to-be-recommended media object first needs to be uploaded through a local device (that is, the first local device, for example, the user terminal 10a). For ease of understanding, for example, the to-be-recommended media object herein is a livestreaming media object, the multimedia data of the to-be-recommended media object is to-be-recommended livestreaming data, and the historical media object is a historical livestreaming media object. Because a livestreaming progress of the to-be-recommended livestreaming data continuously and dynamically changes, in the embodiments of the present disclosure, to ensure reliability and accuracy of mining live content, the to-be-recommended livestreaming data may be dynamically sampled based on a sampling window (for example, a target

sampling window) with a preset window length, to capture to-be-recommended livestreaming data in the target sampling window, and extract, from the captured to-be-recommended livestreaming data, audio data for speech recognition as target audio data. For example, in the embodiments of the present disclosure, audio data of a speaker (for example, voice data or speech data of a streamer object) that is extracted from the to-be-recommended livestreaming data may be used as the target audio data. In the embodiments of the present disclosure, if the target audio data is directly extracted from the to-be-recommended livestreaming data (that is, the multimedia data of the to-be-recommended media object), speech recognition processing may be further performed on the target audio data through a speech conversion understanding component, to obtain target speech recognition text.

[0064] In some embodiments, in the real-time livestreaming scenario, because the audio data of the speaker (for example, the voice data or the speech data of the streamer object) is prone to interference from background noise, colloquial expression of the speaker, repeated speech content, non-uniqueness of the speaker (for example, audio data of another speaker exists), and other factors, it may be difficult to accurately extract target audio data for effective speech recognition, or currently extracted target speech data becomes invalid audio data on which speech recognition can hardly be performed. Based on this, in the embodiments of the present disclosure, when no target audio data is directly extracted from the to-be-recommended livestreaming data, video data may be further quickly extracted from the to-be-recommended livestreaming data based on association information between multi-modal data (for example, different multi-modal data such as voice data and video data) included in the to-be-recommended livestreaming data, and then text information visually shown in the video data may be extracted from the video data, so that the extracted text information is equivalently used as target speech recognition text. Then text vector representation processing is performed on the target speech recognition text in a text vector representation processing mode to obtain a text feature of the target speech recognition text. Then the text feature of the target speech recognition text may be used as a target text representation vector for representing the to-be-recommended media object. Specifically, in the embodiments of the present disclosure, a semantic understanding model may be invoked to perform semantic information extraction on the target speech recognition text based on guiding prompt template data to obtain target semantic understanding information, and text vector representation processing is performed on the target semantic understanding information to obtain the target text representation vector for representing the to-be-recommended media object (for example, the to-be-recommended livestreaming media object, which may be specifically, for example, to-be-recommended livestreaming data, currently obtained through dynamic sampling based on the target sampling

window, of the to-be-recommended livestreaming media object).

[0065] The recommendation method for a media object in the embodiments of the present disclosure specifically includes: obtaining N historical text representation vectors for representing N historical livestreaming media objects, and separately performing vector similarity calculation on the target text representation vector for representing the to-be-recommended livestreaming media object and the N historical text representation vectors of the N historical livestreaming media objects to obtain N vector similarities associated with the N historical livestreaming media objects; then selecting K historical livestreaming media objects from the N historical livestreaming media objects based on the N vector similarities obtained through calculation, and establishing a similarity association relationship between the to-be-recommended livestreaming media object and the K historical livestreaming media objects; and further, obtaining K historical recommendation features of the K historical livestreaming media objects, performing pooling processing on the K historical recommendation features to obtain a target recommendation representation for representing the to-be-recommended livestreaming media object, and then implementing boosted recommendation for the to-be-recommended livestreaming media object based on the target recommendation feature.

[0066] In the embodiments of the present disclosure, a local device (that is, the first local device, for example, the user terminal 10a) for capturing multimedia data of the to-be-recommended livestreaming media object (namely, the to-be-recommended media object) and a local device (that is, the second local device, for example, the user terminal 20a) used by the target service object are different user terminals. Servers for speech recognition processing, video text recognition, text information extraction, text vector representation processing, similarity calculation, and the like in the embodiments of the present disclosure are a same server.

[0067] In some embodiments of the present disclosure, a process of data exchange between the first local device, the cloud server, and the second local device is described by using an example in which the to-be-recommended media object is a livestreaming media object (for example, a current livestreaming room), the multimedia data is to-be-recommended livestreaming data, and the historical media object is a historical livestreaming media object (for example, a historical livestreaming room). For example, when the streamer object starts livestreaming by using a local device (that is, the first local device, for example, the user terminal 10a), the first local device (the user terminal 10) may continuously capture to-be-recommended livestreaming data of the livestreaming media object, and may transmit the continuously captured to-be-recommended livestreaming data to the cloud server (for example, the server 10d in the embodiment corresponding to FIG. 1) through livestreaming. In this way, the cloud server (for example, the

server 10d) may dynamically capture currently received to-be-recommended livestreaming data based on a preset capture window (namely, the target sampling window), so that target audio data can be subsequently extracted from to-be-recommended livestreaming data in the target sampling window. Further, the cloud server may perform speech recognition processing on the target audio data to obtain a speech recognition processing result. If the speech recognition processing result indicates that the target audio data is recognized as including audio data of a speaker (for example, voice data of the streamer object, namely, speech content of the speaker), target speech recognition text obtained by performing text conversion on the recognized audio data of the speaker may be outputted. If the speech recognition processing result indicates that the target audio data is recognized as including no audio data of a speaker, video data needs to be further extracted from the to-be-recommended livestreaming data to recognize text information in the video data, and the recognized text information is equivalently used as target speech recognition text. Further, to ensure accuracy and reliability of semantic understanding and semantic extraction, in the embodiments of the present disclosure, the semantic understanding model of the speech conversion understanding component may be invoked to perform information extraction on the target speech recognition text based on the guiding prompt template data to obtain the target semantic understanding information, so that text vector representation processing is performed on the target semantic understanding information to obtain the target text representation vector for representing the to-be-recommended livestreaming media object. That is, the target text representation vector herein is a text feature obtained by performing semantic understanding on the target speech recognition text. When obtaining the N historical text representation vectors of the N historical livestreaming media objects, the cloud server (for example, the server 10d) separately performs vector similarity calculation on the target text representation vector and the N historical text representation vectors to obtain the N vector similarities, and then selects the K historical livestreaming media objects from the N historical livestreaming media objects based on the N vector similarities to directly establish the similarity association relationship between the to-be-recommended livestreaming media object and the K historical livestreaming media objects. In the embodiments of the present disclosure, an explicit association relationship between each of the K historical livestreaming media objects and the to-be-recommended livestreaming media object (for example, each historical livestreaming media object is highly similar to the to-be-recommended livestreaming media object) can be fully discovered through vector similarity search. Then a similarity association relationship between each historical livestreaming media object and the to-be-recommended livestreaming media object may be directly established based on the discovered explicit association relationship, so that the target recommendation feature for recommending the to-be-recommended livestreaming media object can

be subsequently directly obtained through equivalent calculation based on the K historical recommendation features of the K historical livestreaming media objects. Further, the K historical recommendation features of the K historical livestreaming media objects are obtained, and pooling processing is performed on the obtained K historical recommendation features to obtain the target recommendation feature for representing the to-be-recommended livestreaming media object. Further, the target recommendation model determines, based on the target recommendation feature, a target service object matching the to-be-recommended livestreaming media object, to recommend the to-be-recommended livestreaming media object. In this way, when the target service object is an online user accessing the social software platform through the second local device (for example, the user terminal 20a), the to-be-recommended livestreaming media object recommended by the cloud server through boosted recommendation may be viewed on the social software platform. For example, the second local device may be configured to play the to-be-recommended livestreaming data of the to-be-recommended livestreaming media object through the social software platform.

[0068] The to-be-recommended media object and the historical media object are data of a same type. For example, if the multimedia data of the to-be-recommended media object is livestreaming data, multimedia data of the historical media object is also livestreaming data; or if the multimedia data of the to-be-recommended media object is video-on-demand data, multimedia data of the historical media object is also video-on-demand data. There are more cases. Details are not described herein again.

[0069] In the embodiments of the present disclosure, a service terminal corresponding to the to-be-recommended media object (for example, the to-be-recommended livestreaming media object) and a service terminal corresponding to a target matching user may include intelligent terminals such as a smartphone, a tablet computer, a notebook computer, and a desktop computer.

[0070] In the embodiments of the present disclosure, a computer device for processing the multimedia data of the to-be-recommended media object may be a server (for example, the cloud server), or may be any user terminal (for example, the first local device) in the user terminal cluster. The server may be an independent physical server, or may be a server cluster or a distributed system that includes a plurality of physical servers, or may be a cloud server that provides basic cloud computing services such as a cloud service, a cloud database, cloud computing, a cloud function, cloud storage, a network service, cloud communication, a middleware service, a domain name service, a security service, a CDN, big data, and an artificial intelligence platform.

[0071] Further, FIG. 2 is a diagram of a data exchange scenario according to an embodiment of the present disclosure. A user terminal 10a shown in FIG. 2 may be configured to capture

multimedia data (for example, livestreaming data displayed on a livestreaming interface of a streamer A shown in FIG. 2) of a media object (that is, a to-be-recommended media object, where the to-be-recommended media object may be a to-be-processed media object shown in FIG. 2, for example, a livestreaming room to which the streamer A shown in FIG. 2 belongs) that needs boosted recommendation, to transmit the captured multimedia data to a server 10d shown in FIG. 2. The livestreaming interface of the streamer A is a display interface of a livestreaming room to which a streamer object (for example, the streamer A) belongs.

[0072] In a real-time livestreaming scenario, with proceeding of a livestreaming progress of the livestreaming room to which the streamer A belongs, an amount of data information of the multimedia data (namely, the livestreaming data) captured in the livestreaming room to which the streamer A belongs may be huge.

[0073] Therefore, in one embodiment, to help the server 10d quickly implement in-depth understanding of the multimedia data in the livestreaming room (to be specific, livestreaming data represented by a video and/or audio in the livestreaming room), in the embodiments of the present disclosure, when capturing the multimedia data (namely, the livestreaming data), the user terminal 10a may capture a segment of live video (in other words, capture a video clip) from the multimedia data in the livestreaming room based on a preset sampling window (namely, a target sampling window) as a to-be-recommended livestreaming data segment in the target sampling window, so that the to-be-recommended livestreaming data segment is collectively referred to as dynamically captured to-be-recommended multimedia data. Then all to-be-recommended livestreaming data segments (namely, to-be-recommended multimedia data) dynamically captured by the user terminal 10a may be transmitted to the server 10d. In this way, the server 10d can discover, through key frame extraction, optical character recognition (OCR), automatic speech recognition (ASR), or the like, text information of the livestreaming room from the received to-be-recommended livestreaming data segments (namely, the to-be-recommended multimedia data), so that a target text representation vector for representing the livestreaming room can be dynamically determined based on currently newly discovered text information of the to-be-recommended livestreaming data segments (namely, the to-be-recommended multimedia data).

[0074] For example, when capturing the multimedia data (namely, the livestreaming data) in the livestreaming room, the user terminal 10a may first obtain, based on the livestreaming progress (namely, media data association duration) of the livestreaming room, a target dynamic sampling time point, window duration, and a sampling interval duration that are related to the livestreaming progress (namely, the media data association duration). The target dynamic sampling time point is a

time point changing with a change in the livestreaming progress (namely, the media data association duration) of the to-be-recommended media object (that is, a livestreaming media object, for example, the livestreaming room). For example, in the embodiments of the present disclosure, a time point (that is, the 1st dynamic sampling time point) at which the livestreaming progress (namely, the media data association duration, to be specific, start duration of the to-be-recommended livestreaming data) is the window duration may be referred to as a first dynamic sampling time point. Similarly, with a change in the livestreaming progress, in the embodiments of the present disclosure, another time point (that is, the 2nd dynamic sampling time point) at which a new livestreaming progress (namely, new media data association duration) is the window duration may be referred to as a second dynamic sampling time point. The second dynamic sampling time point herein is a next time point of the first dynamic sampling time point. Duration between the first dynamic sampling time point and the second dynamic sampling time point is the sampling interval duration. By analogy, a next time point of the second dynamic sampling time point may be used as a third dynamic sampling time point. In this way, the user terminal 10a may collectively refer to, as the target dynamic sampling time point, different dynamic sampling time points (for example, the first dynamic sampling time point, the second dynamic sampling time point, and the third dynamic sampling time point) that are sequentially determined with a change in the livestreaming progress, and collectively refer to, as the target sampling window, sampling windows in which different dynamic sampling time points are used as data sampling end time points; and then may transmit, to the server 10d, the to-be-recommended livestreaming data segment captured based on the target sampling window as to-be-recommended livestreaming data.

[0075] Specifically, for the first dynamic sampling time point, a sampling window in which the first dynamic sampling time point is used as a data sampling end time point may be referred to as a first sampling window. Similarly, another sampling window in which the second dynamic sampling time point is used as a data sampling end time point may also be referred to as a second sampling window. When duration between the first dynamic sampling time point and a next time point of the first dynamic sampling time point is the sampling interval duration, the second dynamic sampling time point is determined, and then the first dynamic sampling time point is replaced with the second dynamic sampling time point used as an updated target dynamic sampling time point. Then the second sampling window may be used as the target sampling window, to capture new to-be-recommended livestreaming data (namely, a new to-be-recommended livestreaming data segment), to achieve dynamic sampling. Details are not described herein again. The sampling interval duration is less than or equal to the window duration, so that multimedia data (namely, livestreaming data) in the entire livestreaming room can be fully captured. In addition, window

duration of the first sampling window is the same as window duration of the second sampling window.

[0076] In some embodiments, to avoid occupation of additional storage space due to a huge amount of data information captured by the user terminal 10a, in the embodiments of the present disclosure, when the multimedia data (for example, the livestreaming data) in the livestreaming room is captured, the multimedia data (for example, the livestreaming data) in the livestreaming room may be further directly transmitted to the server 10d through data streaming, so that the server 10d can dynamically determine, based on a strong computing capability of the server, different dynamic sampling time points on the server side in the foregoing manner of determining the target dynamic sampling time point by the user terminal 10a, and sampling windows in which different dynamic sampling time points are used as data sampling end time points are collectively referred to as the target sampling window. Then a to-be-recommended livestreaming data segment captured based on the target sampling window may be used as to-be-recommended livestreaming data, to subsequently implement real-time boosted recommendation for the livestreaming room with a small interaction data amount in a recommendation system through ASR and large language model (LLM) integration, and achieve timeliness and precision of data distribution of the livestreaming data in the livestreaming room. In other words, in the present disclosure, in a real-time video/audio recommendation scenario represented by livestreaming, more time-effective content understanding may be performed on a dynamically captured local live video clip (to be specific, a live video clip captured based on the target sampling window, for example, the to-be-recommended livestreaming data segment) in the livestreaming room through ASR and LLM integration, so that K historical livestreaming rooms (specifically, local live video clips captured based on a historical sampling window) highly similar to the livestreaming room (which may be specifically a local live video clip in the target sampling window) can be subsequently determined from N historical livestreaming rooms (for example, popular livestreaming rooms) that have larger interaction data amounts than the livestreaming room (for example, a less popular livestreaming room), to establish an explicit association relationship. Then a target recommendation feature for recommending the livestreaming room may be obtained through equivalent calculation based on K historical recommendation features that are recorded in the recommendation system and that are historically configured for recommending the K historical livestreaming rooms (to be specific, popular livestreaming rooms that have large interaction data amounts and therefore have sufficient interaction features), to alleviate, by using the target recommendation feature, a problem that the livestreaming room (for example, the less popular livestreaming room) has sparse interaction features due to a small volume of interaction data with a user.

[0077] As shown in FIG. 2, when receiving the multimedia data (for example, the to-be-recommended livestreaming data segment), transmitted by the user terminal 10a, of the to-be-processed media object (namely, the to-be-recommended media object), the server 10d obtains target audio data in the multimedia data through operation S11. To be specific, in the embodiments of the present disclosure, when the to-be-recommended livestreaming data segment is pre-captured or captured in real time based on the target sampling window, the server 10d may further extract audio data from the to-be-recommended livestreaming data segment as the target audio data. Further, the server 10d obtains a speech conversion understanding component and a similarity search component for performing speech conversion processing on the target audio data. Then the server 10d performs operation S12 to perform speech recognition processing through a speech recognizer, to be specific, perform speech recognition processing on the target audio data through the speech recognizer in the speech conversion understanding component, to obtain a speech recognition processing result. If the speech recognition processing result is that no target speech recognition text of the target audio data is recognized (to be specific, if the speech recognition processing result is that speech text corresponding to voice data of a speaker in the target audio data is recognized, the recognized speech text is used as target speech recognition text), the server 10d performs operation S13 to perform text semantic understanding processing through a semantic understanding model. Specifically, text semantic understanding processing is performed based on guiding prompt template data through the semantic understanding model in the speech conversion understanding component, to obtain target semantic understanding information associated with the target speech recognition text. Then the server 10d performs operation S14 to perform text vector representation processing through a text representer. To be specific, text vector representation processing is performed on the target semantic understanding information through the text representer in the similarity search component (to be specific, the target semantic understanding information is vectorized to obtain a text feature of the target speech recognition text), to obtain a target text representation vector for representing the to-be-recommended media object.

[0078] Specifically, if a type of the to-be-recommended media object is a video-on-demand type, prompt template data associated with the video-on-demand type is used as the guiding prompt template data. In addition, if the speech recognition processing result is that no target speech recognition text of the target audio data is recognized (to be specific, the speech recognition processing result is that no speech text corresponding to the voice data of the speaker in the target audio data is recognized), the target audio data is marked as invalid audio data. Then video data in the multimedia data (for example, the to-be-recommended livestreaming data segment), captured in the target sampling window, of the to-be-recommended media object is obtained, to recognize text

in an image (namely, a video frame) in the video data, and the text is used as target speech recognition text.

[0079] Further, the server 10d performs operation S21 to obtain a historical text representation vector. Specifically, the server 10d may obtain N historical text representation vectors of N historical media objects (for example, N historical livestreaming media objects) from a preset historical media object database, and then calculate vector similarities through operation S31, to be specific, separately calculate a vector similarity between the target text representation vector and each of the N historical text representation vectors to obtain N vector similarities. As shown in FIG. 2, the server 10d performs operation S32 to select top K historical media objects (for example, K historical livestreaming media objects) with high vector similarities from the N historical media objects (for example, the N historical livestreaming media objects). Specifically, K historical text representation vectors with similarities ranked top are selected from the N vector similarities, so that historical livestreaming media objects respectively corresponding to the K historical text representation vectors can be selected, to obtain the K historical livestreaming media objects. Then a similarity association relationship between a to-be-recommended livestreaming media object (namely, the to-be-recommended media object) and the K historical livestreaming media objects may be established. Further, the server 10d performs operation S33 to obtain historical recommendation features of the K historical media objects, and then performs operation S34 to calculate a target recommendation feature based on the K historical recommendation features. The target recommendation feature is a recommendation feature that needs to be used by a recommendation system provided by the server 10d to perform media recommendation (for example, boosted recommendation) for the to-be-recommended livestreaming media object. Further, the server 10d may perform operation S35 to recommend the to-be-recommended media object to a target service object based on the target recommendation feature. To be specific, when the server 10d obtains, through calculation, the target recommendation feature for performing boosted recommendation for the to-be-recommended media object, the recommendation system may further invoke a target recommendation model, so that the target recommendation model can find, through targeted matching, a user (namely, the target service object) matching the to-be-recommended livestreaming media object (namely, the to-be-recommended media object) from a large quantity of users based on the target recommendation feature, to recommend the to-be-recommended livestreaming media object to the determined target service object. This can not only improve the boosted recommendation for the to-be-recommended media object, but also increase precision of the boosted recommendation.

[0080] The N historical livestreaming media objects are media objects whose interaction

data amounts reach a threshold and that are extracted from all historical livestreaming data based on an interaction data curve distribution corresponding to the historical media object database. Then historical text representation vectors of the N historical livestreaming media objects may be obtained from a historical text vector representation database corresponding to the historical media object database, to obtain the N historical text representation vectors. The N historical text representation vectors are obtained by processing historical livestreaming data of the N historical livestreaming media objects, and a specific processing process of processing, by the server 10d, the historical livestreaming data of the N historical livestreaming media objects is the same as the specific processing process of processing, by the server 10d, the to-be-recommended livestreaming data of the to-be-recommended livestreaming media object. Details are not described herein again. Historical audio data corresponding to any historical text representation vector is audio data recorded (for example, last recorded) by the server 10d for a corresponding historical media object in a historical sampling window. To be specific, the server 10d may obtain target audio data of a corresponding historical media object in a historical sampling window in which a historical dynamic sampling time point is an end historical dynamic sampling time point recorded when playing of historical livestreaming data ends and window duration is historical window duration. The historical window duration is the same as window duration of the target sampling window.

[0081] To enable the target recommendation model to recommend, based on the target recommendation feature, the to-be-recommended livestreaming media object to the target service object matching the to-be-recommended livestreaming media object, first, N pieces of historical interaction behavior information of the N historical livestreaming media objects need to be obtained, and the N pieces of historical interaction behavior information are used as a training sample for training an initial recommendation model. Further, the server 10d obtains a recommendation indicator of the training sample, and performs model training on the initial recommendation model by using the recommendation indicator and the training sample to obtain the target recommendation model.

[0082] Further, when the target recommendation model is deployed on a service processing platform (for example, a social software platform) of the user terminal 20a, the target recommendation model may be invoked to recommend the to-be-recommended livestreaming media object to a target service object matching to-be-recommended media data (for example, to-be-recommended livestreaming data corresponding to the display interface of the streamer A shown in FIG. 2) of the to-be-recommended livestreaming media object (for example, the livestreaming room to which the streamer A shown in FIG. 2 belongs). The target service object is an online user accessing the social software platform on the user terminal 20a. Then the server 10d transmits the

to-be-recommended livestreaming data to the user terminal 20a, so that the user terminal 20a displays the to-be-recommended livestreaming media object in a recommendation display list W when receiving the to-be-recommended livestreaming data of the to-be-recommended livestreaming media object.

[0083] Specifically, when the target recommendation model in the recommendation system is deployed on a service processing platform (for example, a social software platform), the server 10d may obtain object association data information of a service object associated with the service processing platform (for example, the social software platform), and input the object association data information to the target recommendation model, to perform information vector representation extraction through the target recommendation model to obtain an object vector representation feature of the service object. In this way, the target recommendation model can select, from the object vector representation feature of the service object, an object vector representation feature matching the target recommendation feature, and use a service object corresponding to the selected object vector representation feature as the target service object matching the to-be-recommended livestreaming media object. Further, the target recommendation model may use the to-be-recommended livestreaming media object as a target recommendation object to be recommended to the target service object, and push the target recommendation object to a terminal device (for example, the user terminal 20a shown in FIG. 2) corresponding to the target service object, so that the terminal device (for example, the user terminal 20a shown in FIG. 2) displays the target recommendation object in a recommendation display list (for example, the recommendation display list W shown in FIG. 2) of the service processing platform.

[0084] Therefore, in the embodiments of the present disclosure, the K historical media objects associated with the to-be-recommended media object may be accurately selected by associating the target text representation vector of the to-be-recommended media object with the historical text representation vectors of the historical media objects. Further, in the embodiments of the present disclosure, when the historical recommendation features of the K historical media objects are obtained, the target recommendation feature of the to-be-recommended media object may be directly obtained through processing (to be specific, obtained through equivalent calculation) based on the historical recommendation features of the K historical media objects, so that the target recommendation model can be used (namely, invoked) to recommend, based on the target recommendation feature, the to-be-recommended media object to the target service object matching the to-be-recommended media object.

[0085] Further, FIG. 3 is a schematic flowchart of a recommendation method for a media

object according to an embodiment of the present disclosure. The method may be performed by a server, or may be performed by the foregoing first local device and a server through interaction. For example, when capturing multimedia data of a to-be-recommended media object, the first local device (for example, the user terminal 10a corresponding to FIG. 1) transmits the multimedia data to the server (for example, the server 10d in the embodiment corresponding to FIG. 1), so that the server can perform data processing on the received multimedia data according to the recommendation method for a media object in this embodiment of the present disclosure, to finally find, through targeted matching, a target service object matching the to-be-recommended media object. In this way, the server can subsequently accurately transmit the multimedia data of the to-be-recommended media object to the target service object. To be specific, the to-be-recommended media object may be displayed in a recommendation display list on a service processing platform (for example, a social software platform) of a second local device (for example, the user terminal 20a corresponding to FIG. 1) used by the target service object. In this embodiment of the present disclosure, the recommendation method for a media object may include at least the following operation S101 to operation S104:

[0086] Operation S101: Obtain target audio data of a to-be-recommended media object in a target sampling window, and perform text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the to-be-recommended media object.

[0087] The target audio data is audio data, captured based on the target sampling window, of the to-be-recommended media object. The target text representation vector is obtained by performing text vector representation processing on the target speech recognition text associated with the target audio data.

[0088] A specific process of obtaining, by the server, the target text representation vector for representing the to-be-recommended media object may be described as follows: obtaining the target audio data of the to-be-recommended media object in the target sampling window, and then obtaining a speech conversion understanding component and a similarity search component for performing speech conversion processing on the target audio data. The speech conversion understanding component includes a speech recognizer and a semantic understanding model for performing semantic extraction, and the similarity search component includes a text representer. Further, the server performs speech recognition processing on the target audio data through the speech recognizer (to be specific, the speech recognizer determines whether voice data or speech data of a speaker exists in the target audio data, and if voice data or speech data of a speaker exists,

may recognize the voice data or the speech data of the speaker as the target speech recognition text) to obtain a speech recognition processing result. If the speech recognition processing result indicates that the target speech recognition text of the target audio data is recognized, text semantic understanding processing is performed on the target speech recognition text through the semantic understanding model to obtain target semantic understanding information of the target speech recognition text. Vector representation is performed on the target semantic understanding information through the text representer to obtain the target text representation vector for representing the to-be-recommended media object.

[0089] In some embodiments, in an implementation, if the speech recognition processing result indicates that no target speech recognition text of the target audio data is recognized, video data of the to-be-recommended media object is obtained from multimedia data of the to-be-recommended media object based on the target sampling window, and the obtained video data is used as target video data of the to-be-recommended media object in the target sampling window. Video content recognition is performed on a video frame in the target video data to obtain video text content for representing the target video data, and the target speech recognition text is updated by using the video text content.

[0090] In some embodiments, in another implementation, if the speech recognition processing result indicates that no target speech recognition text of the target audio data is recognized, the target audio data is marked as invalid audio data. The invalid audio data herein is target speech data extracted from the multimedia data of the to-be-recommended media object based on the target sampling window, and is essentially audio data determined as including no voice data of a speaker. In this case, the server may discard target speech data that is currently marked as invalid audio data, to avoid a waste of storage resources and computing resources on the server.

[0091] For ease of understanding, herein, for example, the to-be-recommended media object is a livestreaming media object, and the multimedia data (for example, a to-be-recommended livestreaming data segment), obtained through sampling based on the target sampling window, of the to-be-recommended media object is to-be-recommended livestreaming data. In this case, a historical media object is a historical livestreaming media object, and multimedia data of the historical media object is historical livestreaming data. In a real-time livestreaming scenario, there are two cases: The to-be-recommended livestreaming data captured based on the target sampling window includes audio data (to be specific, audio data related to voice of a speaker exists) or includes no audio data (to be specific, no audio data related to voice of a speaker exists). Therefore, when audio data (that is, the target audio data) in the target sampling window is extracted from the

to-be-recommended livestreaming data, speech recognition processing may be performed on the target audio data through the speech recognizer in the speech conversion understanding component to obtain a speech recognition processing result. Further, if the speech recognition processing result indicates that audio data exists (to be specific, audio data related to voice of a speaker exists) in the to-be-recommended livestreaming data, speech conversion may be performed on the audio data related to the voice of the speaker that exists in the to-be-recommended livestreaming data, and speech recognition text obtained through speech conversion is used as target speech recognition text that is outputted by the speech recognizer and that is associated with the target audio data of the to-be-recommended livestreaming data. If the speech recognition processing result indicates that the to-be-recommended livestreaming data includes no audio data (to be specific, no audio data related to voice of a speaker exists), the target audio data is marked as invalid audio data, and the invalid audio data is not to be used subsequently. Then video data of the to-be-recommended livestreaming data may be obtained as target video data, to recognize text information in the target video data. The recognized text information in the target video data is used as video recognition text (namely, video text content) that is outputted by the speech recognizer and that is associated with the to-be-recommended livestreaming data, so that the video recognition text (namely, the video text content) can be equivalently used as target speech recognition text.

[0092] When the speech recognizer performs speech recognition by using an ASR algorithm, for a specific manner of performing speech recognition by using the ASR algorithm to obtain the target speech recognition text, refer to the following formula (1):

$$\text{text}_i = \text{asr}(i_c) \quad (1)$$

[0093] i represents the to-be-recommended media object, for example, a to-be-boosted cold-start/long-tail livestreaming room, referred to as a current livestreaming room for short. c represents a target sampling window at a current timestamp (to be specific, a target dynamic sampling time point that is newly determined based on a livestreaming progress of the current livestreaming room). Therefore, i_c is configured for representing to-be-recommended livestreaming data, in the target sampling window, that is obtained, through sampling, from multimedia data of the current livestreaming room based on the target sampling window. asr represents the ASR algorithm. In this case, the speech recognizer may extract target speech data from the to-be-recommended livestreaming data in the target sampling window by using the ASR algorithm, to perform speech recognition on audio data that is related to voice of a speaker and that exists in the target speech data, to obtain ASR text information (namely, text_i) for representing the current livestreaming room (namely, i).

[0094] Further, a specific process of performing, by the server through the semantic understanding model for semantic extraction in the speech conversion understanding component, semantic understanding processing on the target speech recognition text obtained through speech recognition, to obtain the target semantic understanding information may be described as follows: obtaining guiding prompt template data for guiding the semantic understanding model to perform model output; inputting the guiding prompt template data and the target speech recognition text to the semantic understanding model, and performing semantic information extraction on the target speech recognition text through the semantic understanding model based on data content in the guiding prompt template data to obtain target semantic extraction information of the target speech recognition text; and further, determining the target semantic understanding information of the target speech recognition text based on the target semantic extraction information.

[0095] When the semantic understanding model performs semantic extraction, the semantic understanding model needs to be guided by a template, to output the target semantic understanding information. For example, when the semantic understanding model obtained by the server (for example, the server 10d in the embodiment corresponding to FIG. 1) is an LLM, the LLM may have an understanding capability and an inference capability based on currently obtained guiding prompt template data. For example, the LLM may implement semantic understanding and inference analysis on the target speech recognition text based on the guiding prompt template. For example, when an object type of the to-be-recommended media object is a video livestreaming type (to be specific, the to-be-recommended media object is the foregoing to-be-recommended livestreaming media object), the guiding prompt template data may specifically include prompt template data associated with the video livestreaming type, for example, a livestreaming media object style (for example, a streamer style), a livestreaming theme, live content, a community atmosphere, a livestreaming room viewer, and a potential topic. This is not limited herein.

[0096] Further, the server inputs the guiding prompt template data and the target speech recognition text to the LLM, and the LLM performs semantic information extraction on the target speech recognition text based on the guiding prompt template data to obtain the target semantic extraction information associated with the target speech recognition text. Then the LLM outputs target semantic understanding information including the following themes: the streamer style, the livestreaming theme, the live content, the community atmosphere, the livestreaming room viewer, the potential topic, and the like. Specific output content of the LLM is not limited herein.

[0097] When the semantic understanding model is the LLM, a process of performing semantic understanding on the target speech recognition text is as follows: Specifically, the target

semantic understanding information may be determined by using the following formula (2):

$$\text{semantic}_i = \text{llm}\big(\text{promt}(\text{text}_i)\big) \qquad (2)$$

[0098] text_i represents the target speech recognition text obtained through recognition by using the formula (1). i represents a to-be-boosted cold-start/long-tail livestreaming room. prompt represents an LLM text generation template (namely, the guiding prompt template data). llm represents a pre-trained LLM. The ASR text information (namely, the target speech recognition text) and generated multi-dimensional semantic information (namely, semantic_i) of the livestreaming room may be classified as natural language. Therefore, in this embodiment of the present disclosure, inference analysis may be performed on the target speech recognition text based on the guiding prompt template data by using the inference capability of the LLM, so that multi-dimensional semantic information extraction can be performed on an inference analysis result obtained through inference analysis. Then extracted multi-dimensional semantic information may be collectively referred to as the target semantic extraction information of the target speech recognition text.

[0099] If an object type of the to-be-recommended media object is a video livestreaming type, prompt template data associated with the video livestreaming type is used as the guiding prompt template data. In some embodiments, if an object type of the to-be-recommended media object is a video-on-demand type, prompt template data associated with the video-on-demand type is used as the guiding prompt template data.

[00100] When the multimedia data of the to-be-recommended media object is to-be-recommended livestreaming data, the guiding prompt template data is the prompt template data associated with the video livestreaming type. Similarly, if the multimedia data of the to-be-recommended media object is video-on-demand data, the guiding prompt template data is the prompt template data associated with the video-on-demand type, for example, a theme such as character lines, a character style, or a scene style. This is not limited herein.

[00101] Further, in this embodiment of the present disclosure, vector representation (that is, text vector representation processing) may be performed, through the text representer, on the target semantic understanding information outputted by the semantic understanding model, to obtain the target text representation vector for subsequent similarity search.

[00102] Operation S102: Obtain N historical text representation vectors for representing N historical media objects, and separately perform vector similarity calculation on the target text representation vector and each historical text representation vector to obtain N vector similarities.

[00103] Each of the N historical text representation vectors is obtained by performing speech conversion processing on historical audio data of each historical media object through the speech conversion understanding component and the similarity search component. The N vector similarities herein are obtained by separately performing vector similarity calculation on the target text representation vector and each of the N historical text representation vectors.

[00104] The target text representation vector is obtained by performing speech conversion processing on the target audio data through the speech conversion understanding component and the similarity search component for speech conversion processing.

[00105] A specific process of obtaining, by the server, the N historical text representation vectors for representing the N historical media objects includes: obtaining interaction data curve distribution corresponding to a historical media object database, the interaction data curve distribution being determined based on each historical media object in the historical media object database and an interaction data amount of the historical media object; further, obtaining, from the historical media object database based on the interaction data curve distribution, the N historical media objects with interaction data amounts reaching an interaction data threshold; and then obtaining the N historical text representation vectors of the N historical media objects from a historical text vector representation database associated with the N historical media objects. Each of the N historical text representation vectors is obtained by performing speech conversion processing on historical audio data of each historical media object through the speech conversion understanding component and the similarity search component. In addition, each piece of historical audio data is audio data of each of the N historical media objects in a recorded historical sampling window.

[00106] For ease of understanding, FIG. 4 is a diagram of an interaction data curve of historical media objects according to an embodiment of the present disclosure. A historical database obtained by the server may specifically include the historical media object database and the historical text vector representation database. In this case, the server may statistically collect, based on all historical media objects (for example, historical livestreaming rooms) recorded in the historical media object database, interaction data amounts of the historical media objects (for example, the historical livestreaming rooms), and then may obtain the interaction data curve distribution in FIG. 4 based on values of the statistically collected interaction data amounts. In this way, the server may classify, based on the interaction data threshold configured for the historical media object database, the historical media objects into media objects located in a head region (for example, media objects with interaction data amounts reaching the interaction data threshold, for

example, media objects with interaction data amounts ranked top 10% in FIG. 4) and media objects located in a tail region (for example, media objects with interaction data amounts not reaching the interaction data threshold, for example, media objects with interaction data amounts ranked the last 90% in FIG. 4).

[00107] For example, as shown in FIG. 4, the media objects located in the head region may be historical media objects with interaction data amounts ranked top 10% in the interaction data curve distribution. Then the historical media objects with interaction data amounts ranked top 10% may be collectively referred to as the N historical media objects that need to participate in a recommendation process of boosted recommendation in this embodiment of the present disclosure. Because a type of the multimedia data of the to-be-recommended media object is livestreaming data, a type of the multimedia data of the historical media object is also livestreaming data. Then the server may obtain the N historical text representation vectors of the selected N historical media objects from the historical text vector representation database, and for example, may specifically obtain N pieces of historical livestreaming data of the N historical media objects in the historical sampling window, so that historical text representation vectors respectively corresponding to the N pieces of historical livestreaming data can be quickly selected subsequently, to finally obtain the N historical text representation vectors.

[00108] A process of generating the N historical text representation vectors of the N historical livestreaming media objects is the same as the process of generating the target text representation vector of the to-be-recommended media object. The target audio data is audio data whose duration is window duration and that is dynamically captured in the target sampling window when the target dynamic sampling time point changes with playing of the to-be-recommended livestreaming data (to be specific, changes with a change in media data association duration of the to-be-recommended media object). To be specific, with a dynamic and continuous change in the target dynamic sampling time point, audio data that is dynamically captured by the server and that is within the target sampling window also continuously changes. Therefore, when audio data of a speaker that is captured by the server in a sampling window (for example, a second sampling window) corresponding to a current dynamic sampling time point and audio data of a speaker that is captured by the server in another sampling window (for example, a first sampling window) corresponding to a previous dynamic sampling time point are same audio data, it can be considered that the audio data currently captured in the second sampling window is duplicate with the audio data captured in the first sampling window. Therefore, to avoid a waste of storage resources and computing resources of the server due to capturing of duplicate audio data (for example, captured audio data may be duplicate due to duplicate speech content of a streamer object), the audio data

currently captured in the second sampling window may be collectively referred to as the foregoing invalid audio data.

[00109] For any one of the N historical media objects determined based on the interaction data curve distribution shown in FIG. 4, historical audio data corresponding to the historical livestreaming data may be specifically audio data captured based on the historical sampling window (for example, the historical sampling window may be a recorded sampling window in which sampling duration is window duration and a historical dynamic sampling time point is a time point at which the historical livestreaming data ends). Then speech conversion processing is performed on the historical audio data through the speech conversion understanding component and the similarity search component to obtain the historical text representation vector. Window duration of the target sampling window is the same as window duration of the historical sampling window.

[00110] Specifically, when the semantic understanding model is the LLM, an embodiment of separately performing semantic extraction on the to-be-recommended livestreaming data and the historical livestreaming data may include the following: Specifically, the target semantic understanding information associated with the target speech recognition text may be determined by using the following formula (3), and the historical text representation vector associated with the historical semantic understanding information may be determined by using the following formula (4):

$$\mathrm{semantic_i} = \mathrm{llm}\big(\mathrm{promt}(\mathrm{text_{ci}})\big) \mathrm{i} \in \mathrm{C} \qquad (3)$$

$$\mathrm{semantic_j} = \mathrm{llm}\Big(\mathrm{promt}\big(\mathrm{text_{cj}}\big)\Big) \mathrm{j} \in \mathrm{H} \qquad (4)$$

[00111] $\mathrm{text_{ci}}$ represents target speech recognition text, in a target time window, that is obtained through recognition based on the formula (1). Similarly, $\mathrm{text_{cj}}$ is recorded historical speech recognition text, in a historical time window, that is historically obtained through recognition based on the formula (1). i represents the to-be-recommended media object (for example, a current media object i may be a to-be-boosted current livestreaming room, for example, a cold-start/long-tail livestreaming room). j represents the historical media object (for example, a historical high-interaction livestreaming room, the historical high-interaction livestreaming room herein being a historical livestreaming room in the head region). ci represents a target sampling window at a current timestamp (to be specific, a target dynamic sampling time point that is newly determined based on a livestreaming progress of the current livestreaming room). cj represents a historical sampling window at a historical timestamp (to be specific, a historical dynamic sampling time point that is newly determined based on a livestreaming progress of the historical

livestreaming room, for example, a time point at which livestreaming in the historical livestreaming room ends is used as an end historical dynamic sampling time point). Because window duration of the historical sampling window is consistent with that of the target sampling window, ci=cj=c. That is, c herein may be collectively referred to as a sampling window. prompt represents the LLM text generation template in the formula (2), namely, the guiding prompt template data. llm represents a pre-trained LLM. C represents a sum of all to-be-recommended media objects that currently need to be boosted (that is, a to-be-boosted media object list including all to-be-boosted to-be-recommended media objects). H represents a sum of all of the N historical media objects (that is, a historical media object list including the N historical media objects).

[00112] Further, in one embodiment of the present disclosure, the separately performing text vector representation processing on the target semantic understanding information of the to-be-recommended media object and the historical semantic understanding information of the historical media object may include the following: Specifically, the target text representation vector associated with the target semantic understanding information may be determined by using the following formula (5), the historical text representation vector associated with the historical semantic understanding information may be determined by using the following formula (6), and the vector similarity between the target text representation vector and the historical text representation vector may be obtained by using the following formula (7):

$$e_i = \text{embedding}(\text{semantic}_i) i \in C \quad (5)$$

$$e_j = \text{embedding}(\text{semantic}_j) j \in H \quad (6)$$

$$s_{i,j} = \text{similarity}(e_i, e_j) \quad (7)$$

[00113] i represents any current livestreaming room (for example, a cold-start/long-tail livestreaming room) among all to-be-boosted livestreaming rooms (namely, C). j represents any one of N historical high-interaction livestreaming rooms (namely, H).semantic_i represents the target semantic understanding information in the formula (3). semantic_j represents the historical semantic understanding information in the formula (4). embedding() represents a text vectorization algorithm. similarity() represents a vector similarity metric function. For example, $\text{similarity}(e_i, e_j)$ may indicate to perform vector similarity calculation on the target text representation vector (namely, e_i) and a historical text representation vector (namely, e_j) to obtain a vector similarity (namely, $s_{i,j}$) between the current livestreaming room and a historical high-interaction livestreaming room (collectively referred to as the historical livestreaming room).

[00114] Further, when N pieces of historical livestreaming data of the N historical

livestreaming rooms are obtained based on a sampling window (for example, a historical sampling window), N historical text representation vectors of the N pieces of historical livestreaming data may be calculated by using the formula (4) and the formula (6). Then vector similarity calculation may be separately performed, based on the formula (7), on the target text representation vector associated with the to-be-recommended media object (for example, a to-be-recommended livestreaming media object, for example, a current livestreaming room) and the N historical text representation vectors of the N historical media objects (for example, the N historical livestreaming media objects, for example, the N historical livestreaming rooms). A vector similarity calculation method includes a cosine similarity, a Jaccard coefficient similarity, a Euclidean distance, a Manhattan distance, and the like. This is not limited herein.

[00115] Operation S103: Select, based on the N vector similarities, K historical media objects from the N historical media objects, and establish a similarity association relationship between K historical media objects and the to-be-recommended media object.

[00116] Specifically, the server may sort the N historical media objects based on the N vector similarities to obtain a historical media object list associated with the N historical media objects; and obtain, from the historical media object list, the K historical media objects with vector similarities meeting a similarity retrieval condition, and establish the similarity association relationship between the K historical media objects and the to-be-recommended media object. The similarity association relationship is configured for representing that a list formed by the K historical media objects is a similar media object list, and any media object in the similar media object list has an association relationship (that is, the foregoing explicit association relationship) with the to-be-recommended media object.

[00117] For ease of understanding, the server (for example, the server 10d in the embodiment corresponding to FIG. 1) may sort the obtained N historical livestreaming media objects in descending order of the N vector similarities obtained through calculation, to obtain a historical media object list associated with the N historical livestreaming media objects. Further, the server may select the first K historical livestreaming media objects, namely, top K historical livestreaming media objects, from the historical media object list based on a similarity retrieval condition (for example, a search condition of selecting the first K ones), to establish a similarity association relationship between the K historical livestreaming media objects and the to-be-recommended livestreaming media object.

[00118] FIG. 5 is a diagram of a similarity association relationship according to an embodiment of the present disclosure. A to-be-processed media object shown in FIG. 5 is a to-be-

recommended media object that needs boosted recommendation. For ease of understanding, N=3 and K=2 are used as an example herein to calculate a vector similarity between a target text representation vector (to be specific, a text representation vector for representing the to-be-processed media object) and N historical text representation vectors (herein, N=3, that is, three historical text representation vectors 5b, for example, a historical text representation vector A, a historical text representation vector B, and a historical text representation vector C shown in FIG. 5), to obtain three vector similarities 5c (for example, a vector similarity A, a vector similarity B, and a vector similarity C shown in FIG. 5). Then three historical media objects 5a (for example, a historical media object A, a historical media object B, and a historical media object C shown in FIG. 5) (which may be specifically obtained historical livestreaming data of the three historical media objects 5a, that is, three pieces of historical livestreaming data) are sorted in descending order of the three vector similarities 5c obtained through calculation, to obtain a historical media object list associated with the three pieces of historical livestreaming data. Further, in this embodiment of the present disclosure, historical livestreaming data (herein, K=2, that is, two pieces of historical livestreaming data) of the first K historical media objects is filtered based on the historical media object list to obtain two historical media objects associated with the to-be-processed media object (namely, the to-be-recommended media object).

[00119] As shown in FIG. 5, in this embodiment of the present disclosure, after the first two historical media objects (for example, the media object A and the media object C shown in FIG. 5, which may be specifically, for example, historical livestreaming data of the two historical media objects) are selected from the historical media object list, the similarity association relationship shown in FIG. 5 may be directly established with the to-be-recommended media object (for example, to-be-recommended livestreaming data of the to-be-recommended media object). Then a similar media object list may be constructed based on the selected two historical media objects (for example, the media object A and the media object C shown in FIG. 5). The similarity association relationship shown in FIG. 5 may be configured for representing that each historical media object (for example, each historical livestreaming media object) in the similar media object list has an association relationship (that is, an explicit association relationship) with the to-be-recommended media object (for example, the to-be-recommended livestreaming media object).

[00120] In this embodiment of the present disclosure, top K historical media objects (namely, K historical high-interaction media objects) may be found for each to-be-recommended media object based on the N vector similarities obtained through calculation, to form the similar media object list by using the K historical high-interaction media objects. The similar media object list herein may be a semantic proximity list in the following formula (8):

$$r_i \leftarrow [J_1 \dots J_K] \quad (8)$$

[00121] $[J_1 \dots J_K]$ is configured for representing the semantic proximity list including K historical high-interaction media objects found for a current livestreaming room (for example, a livestreaming room i). Subscripts of the symbol J are identity document (ID) indexes of the K historical high-interaction media objects that are found. K represents K historical media objects (for example, K most similar historical high-interaction livestreaming rooms) that are most similar to the current livestreaming room (for example, the livestreaming room i) and that are found through vector similarity search.

[00122] Operation S104: Obtain K historical recommendation features of the K historical media objects, and determine, based on the similarity association relationship and K historical recommendation features, a target recommendation feature for representing the to-be-recommended media object, and perform media recommendation for the to-be-recommended media object based on the target recommendation feature.

[00123] Specifically, when obtaining the K historical recommendation features of the K historical media objects, the server may perform feature pooling processing on the K historical recommendation features based on the similarity association relationship to obtain a pooling processing feature associated with the K historical recommendation features. Further, the server may determine the pooling processing feature as the target recommendation feature for representing the to-be-recommended media object, to implement media recommendation (for example, boosted recommendation) for the to-be-recommended media object based on the target recommendation feature.

[00124] For example, when the K historical media objects are K historical livestreaming media objects, K historical recommendation features of the K historical livestreaming media objects may be obtained from a recommendation feature database associated with the K historical livestreaming media objects, and then pooling processing may be performed on the obtained K historical recommendation features to obtain a target recommendation feature used for representing the to-be-recommended livestreaming media object. Further, in this embodiment of the present disclosure, the server may further invoke a target recommendation model in a recommendation system to determine, based on the target recommendation feature, a target service object matching the to-be-recommended livestreaming media object, to recommend the to-be-recommended livestreaming media object to the target service object.

[00125] In the embodiments of the present disclosure, audio data of a to-be-recommended media object in a target sampling window may be recognized, to obtain target audio data associated

with the to-be-recommended media object. Further, in the embodiments of the present disclosure, speech recognition processing is performed on the target audio data through a speech recognizer in a speech conversion understanding component to obtain target speech recognition text associated with the target audio data. Further, in the embodiments of the present disclosure, information extraction may be performed on the target speech recognition text based on guiding prompt template data through a semantic understanding model in the speech conversion understanding component to obtain target semantic understanding information associated with the target speech recognition text. Then text vector representation processing is performed on the target semantic understanding information to obtain a target text representation vector for representing the to-be-recommended media object. In this way, in the embodiments of the present disclosure, N historical text representation vectors for representing N historical media objects (for example, N historical media objects in a head region may be obtained based on the foregoing interaction data curve distribution) may be obtained, and then vector similarity calculation is separately performed on the target text representation vector and each of the N historical text representation vectors to obtain N vector similarities. Further, in the embodiments of the present disclosure, K historical media objects highly similar to the to-be-recommended media object are selected from the N historical media objects based on the N vector similarities, and then a similarity association relationship between the to-be-recommended media object and the selected K historical media objects is established. In the embodiments of the present disclosure, when K historical recommendation features of the K historical media objects are obtained, feature pooling processing may be further performed on the K historical recommendation features based on the similarity association relationship to equivalently determine a target recommendation feature for representing the to-be-recommended media object, and then media recommendation may be performed, based on the target recommendation feature, for a target service object matching the to-be-recommended media object. In the embodiments of the present disclosure, speech recognition processing, semantic understanding information extraction, and text vector representation processing are sequentially performed on the target audio data to obtain the target text representation vector for representing the target audio data, and similarity calculation is separately performed on the target text representation vector and a historical text representation vector of each historical media object, to finally find the K historical media objects highly similar to the to-be-recommended media object, so that the target recommendation feature for representing the to-be-recommended media object is obtained through equivalent calculation by using the historical recommendation features of the K historical media objects that are found. In other words, in the embodiments of the present disclosure, even if the to-be-recommended media object is a newly introduced cold-start livestreaming room lacking an

interaction data amount, other historical livestreaming rooms with high semantic similarities to the cold-start livestreaming room may be accurately found, and then a target recommendation feature of the cold-start livestreaming room may be obtained through equivalent calculation by using historical recommendation features determined based on historical livestreaming data in the found historical livestreaming rooms (that is, a high-interaction feature for representing the cold-start livestreaming room may be obtained through simulation), so that the target service object corresponding to the to-be-recommended media object can be subsequently accurately found through a recommendation model (that is, the foregoing target recommendation model), to push the to-be-recommended media data to the target service object in a targeted manner. This not only improves accuracy of performing media recommendation (for example, cold-start recommendation) by the target recommendation model for the to-be-recommended media object, but also can improve matching accuracy of performing targeted matching between the to-be-recommended media object and a suitable service object, to recommend the to-be-recommended media object to the service object through boosted recommendation.

[00126] In an embodiment, when a type of a to-be-recommended media object is a livestreaming media object, the to-be-recommended livestreaming media object is used as a target recommendation object to be pushed to a user A matching the livestreaming media object (to be specific, a currently determined target service object that matches the to-be-recommended media object) in a recommendation service (for example, a livestreaming recommendation service). In this way, when the target service object (for example, the user A) matching the livestreaming media object uses another service (for example, an XX service) of a service processing platform (for example, a social software platform), the livestreaming media object or an advertisement for the livestreaming media object may be inserted into the XX service being used by the target service object (for example, the user A), to push the livestreaming media object to the target service object (for example, the user A).

[00127] For ease of understanding, FIG. 6 is a diagram of a scenario of recommending a livestreaming media object on a service processing platform according to an embodiment of the present disclosure. When a target service object 6a (for example, a user A in FIG. 6) uses another service (for example, an XX service in FIG. 6) of a service processing platform 6c, a local device (that is, a second local device, for example, the user terminal 20a in the embodiment corresponding to FIG. 1) used by the target service object 6a may display a target recommendation object (for example, a to-be-recommended livestreaming media object 6b in FIG. 6) to the user A on an XX page corresponding to the XX service, to recommend the to-be-recommended livestreaming media object 6b to the user A.

[00128] Further, in response to a trigger operation performed by the target service object on the to-be-recommended livestreaming media object 6b shown in FIG. 6, the second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1) jumps to display a multimedia data display interface (to be specific, a diagram of a scenario of livestreaming in a livestreaming room in FIG. 7) of the to-be-recommended livestreaming media object. The livestreaming room herein is the to-be-recommended livestreaming media object.

[00129] For ease of understanding, FIG. 7 is a diagram of a scenario of livestreaming in a livestreaming room according to an embodiment of the present disclosure. A multimedia data display interface of a to-be-processed media object shown in FIG. 7, which may be the foregoing to-be-recommended media object (namely, the to-be-recommended livestreaming media object) that needs boosted recommendation, specifically includes an avatar display region 7a for the to-be-processed media object and a display interface interaction region 7b for the to-be-processed media object. When the to-be-recommended livestreaming media object is a livestreaming room, the multimedia data display interface may be the multimedia data display interface of the to-be-recommended livestreaming media object that the second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1) jumps to display in response to the trigger operation performed by the target service object on the to-be-recommended livestreaming media object 6b in FIG. 6 after the target service object clicks/taps a region in which the to-be-recommended livestreaming media object 6b in FIG. 6 is located.

[00130] Correspondingly, in another embodiment, when a type of a to-be-recommended media object is a video-on-demand media object, the video-on-demand media object is used as a target recommendation object to be pushed to a user A matching the video-on-demand media object (to be specific, a currently determined target service object that matches the to-be-recommended media object) in a recommendation service (for example, a livestreaming recommendation service). Then, in this embodiment of the present disclosure, a video-on-demand link corresponding to the video-on-demand media object may be packaged into a recommendation notification message, and the recommendation notification message is transmitted to a terminal device corresponding to the target service object, for example, a second local device, so that the second local device can display the recommendation notification message on a use page of the target service object (for example, the user A) by using a pop-up window, a floating window, or the like. The target service object (for example, the user A) may click/tap the recommendation notification message, to enable the second local device to enter a play interface of the video-on-demand media object.

[00131] For ease of understanding, FIG. 8 is a diagram of a scenario of recommending a

video-on-demand media object on a service processing platform according to an embodiment of the present disclosure. The second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1) may display, on a use page of a target service object 8a (for example, a user A in FIG. 8) (to be specific, a display page of a service processing platform 8c accessed by the user A), a recommendation notification message 8b by using a pop-up window shown in FIG. 8. Then the target service object 8a (for example, the user A on the service processing platform 8c shown in FIG. 8) may click/tap the recommendation notification message to enter a multimedia data display interface (namely, the foregoing play interface) of the video-on-demand media object, so that the video-on-demand media object is recommended to the user A.

[00132] In other words, the second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1) may jump to display the multimedia data display interface of the video-on-demand media object in response to a trigger operation performed by the target service object on the recommendation notification message. The multimedia data display interface of the video-on-demand media object may include a play progress bar 9a shown in FIG. 9.

[00133] For ease of understanding, FIG. 9 is a diagram of a scenario of playing a video based on a video-on-demand identifier according to an embodiment of the present disclosure. The video-on-demand identifier is an identifier for identifying a to-be-recommended video-on-demand media object (for example, the video-on-demand identifier may be specifically a name of a to-be-processed media object shown in FIG. 9). The multimedia data display interface may be an interface that the second local device (for example, the user terminals 20a, 20b, or 20c in the embodiment corresponding to FIG. 1, or the user terminal 20a corresponding to FIG. 2) jumps to display in response to the trigger operation performed by the target service object on the recommendation notification message after the target service object (for example, the user A shown in FIG. 8) clicks/taps the recommendation notification message 8b.

[00134] In an embodiment, after top K historical livestreaming media objects (namely, historical media objects) are obtained through vector similarity search, a target recommendation feature may be obtained through equivalent calculation based on historical recommendation features of top K pieces of historical livestreaming data. The target recommendation feature may be configured for searching for an online matching user (namely, a target service object) matching a to-be-recommended livestreaming media object (namely, a to-be-recommended media object), so that a second local device can subsequently display, on a social software platform (namely, a service processing platform), a recommendation display list for the online matching user found through matching, and then may display the to-be-recommended livestreaming media object (namely, the to-

be-recommended media object) in the recommendation display list.

[00135] For ease of understanding, FIG. 10 is a diagram of a recommendation display list of a service processing platform according to an embodiment of the present disclosure. The recommendation display list includes a streamer display region 100a and a recommendation list region 100b. Further, the recommendation list region 100b includes a media object corresponding to each streamer in the streamer display region 100a (for example, a media object A1 corresponding to a streamer A1, a media object B1 corresponding to a streamer B1, a media object C1 corresponding to a streamer C1, and a media object D1 corresponding to a streamer D1 in FIG. 10). The media object A1 corresponding to the streamer A1 may be a to-be-processed livestreaming media object 100c recommended by a server through boosted recommendation (that is, the foregoing to-be-recommended livestreaming media object).

[00136] In this way, after the target recommendation model selects a target service object matching a to-be-recommended media object, a second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1) may push the received to-be-recommended media object to a target matching user (namely, the target service object), so that the target matching user can view the to-be-recommended media object in a recommendation display list on a social software platform. The recommendation list region 100b includes the to-be-recommended livestreaming media object 100c. In this way, the target matching user may click/tap, on a multimedia data display interface of the to-be-recommended livestreaming media object 100c shown in FIG. 10, a region of the media object A1 in the recommendation list region 100b to jump for displaying.

[00137] In the embodiments of the present disclosure, speech recognition processing is mainly performed on audio data of a to-be-recommended media object (for example, audio data related to voice of a speaker in the to-be-recommended media object), and then recognized speech text may be processed, to subsequently find a target service object matching the to-be-recommended media object. If a speech recognition processing result indicates that audio data related to voice of a speaker exists in target audio data extracted from to-be-recommended livestreaming data, target speech recognition text associated with the target audio data of the to-be-recommended media object is outputted. In some embodiments, if the speech recognition processing result indicates that the target audio data extracted from the to-be-recommended livestreaming data does not include audio data related to voice of a speaker, the target audio data is marked as invalid audio data, and is not to be used subsequently. When the target audio data is marked as invalid audio data, there may be two cases subsequently. In one case, the process directly

ends, and the to-be-recommended media object is not to be further processed. In another case, video data of the to-be-recommended media object may be further obtained as target video data, and text in the target video data is recognized, so that the recognized text is indirectly used as target speech recognition text for processing.

[00138] FIG. 11 is a flowchart of a processing method for invalid target audio according to an embodiment of the present disclosure. The method may be performed by a server. In this way, after obtaining target audio data of a to-be-recommended media object, the server (for example, the server 10d in the embodiment corresponding to the foregoing FIG. 1) performs speech recognition processing on the target audio data through a speech recognizer in a speech conversion understanding component, to recognize whether speech audio (to be specific, audio data related to voice of a speaker) exists in the target audio data. When the target audio data is recognized as including no audio data related to voice of a speaker, the target audio data may be marked as invalid audio data, to directly end the method process. For example, as shown in FIG. 11, the method may include at least operation S2011 to operation S2016.

[00139] Operation S2011: Obtain target audio data.

[00140] Specifically, target audio data of a to-be-recommended media object is obtained, and whether speech audio exists in the target audio data is detected or recognized during speech recognition processing on the target audio data. The speech audio herein may be audio data related to voice of a speaker.

[00141] Operation S2012: Detect whether audio exists in the target audio data.

[00142] Specifically, speech recognition processing is performed on the target audio data through a speech recognizer in a speech conversion understanding component to obtain a speech recognition processing result. Whether audio (for example, speech audio) exists in the target audio data is detected. If speech audio exists, to be specific, when it is determined that audio data related to voice of a speaker is currently detected in the target audio data, operation S2013 is further performed. For a specific process of performing, by the server, speech recognition processing on the target audio data, refer to the descriptions of the specific process of speech recognition processing in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00143] Operation S2013: Recognize text corresponding to the audio in the target audio data.

[00144] Specifically, when it is detected that speech audio exists in target audio, in other words, audio data related to voice of a speaker exists, text corresponding to the speech audio in the target audio data may be further recognized (that is, target speech recognition text of the target

audio data is recognized).

[00145] Operation S2014: Output the target speech recognition text.

[00146] Specifically, when recognizing the text corresponding to the audio in the target audio data, the server may output specific text information of the recognized text as the target speech recognition text. In this way, after this operation is completed, operation S2016 may be further performed.

[00147] In some embodiments, after operation S2012 is performed, a determining result may alternatively be no. To be specific, when it is determined that no audio data related to voice of a speaker exists in the target audio data, operation S2015 is performed.

[00148] Operation S2015: Mark the target audio data as invalid audio data.

[00149] Specifically, the server may mark the target audio data as invalid audio data when detecting that no audio (that is, speech audio, for example, audio data related to voice of a speaker) exists in the target audio (namely, the target audio data).

[00150] Operation S2016: End.

[00151] The server may further perform text semantic understanding processing on the target speech recognition text through a semantic understanding model in the speech conversion understanding component to obtain target semantic understanding information of the target speech recognition text. For a specific process of performing text semantic understanding processing on the target speech recognition text, refer to the descriptions of the specific process of text semantic understanding processing in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00152] Further, FIG. 12 is a flowchart of another processing method for invalid target audio according to an embodiment of the present disclosure. As shown in FIG. 12, the method may include at least operation S2021 to operation S2028.

[00153] Operation S2021: Obtain target audio data.

[00154] For a specific process of obtaining the target audio data by a server, refer to the descriptions of the specific process of obtaining the target audio data in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00155] Operation S2022: Detect whether audio (for example, speech audio) exists in the target audio data.

[00156] Operation S2023: Recognize text corresponding to the audio in the target audio data.

[00157] Specifically, when detecting that audio (for example, speech audio, for example, audio data related to voice of a speaker) exists in the target audio data, the server may recognize text corresponding to the audio in the target audio data (that is, recognize target speech recognition text of the target audio data). In this way, after completing this operation, the server may further perform operation S2027.

[00158] In some embodiments, as shown in FIG. 12, after performing operation S2022, the server may further perform operation S2024 when a determining result is no.

[00159] Operation S2024: Mark the target audio data as invalid audio data.

[00160] Specifically, when detecting that no audio exists in target audio (for example, no speech audio exists, for example, no audio data related to voice of a speaker exists), the server marks the target audio data as invalid audio data, so that operation S2025 can be further performed subsequently.

[00161] Operation S2025: Obtain video data corresponding to the target audio data.

[00162] Specifically, the server may obtain the video data corresponding to the target audio data of a to-be-recommended media object, and use the video data as target video data. For a specific process of obtaining the target video data by the server, refer to the descriptions of the specific process of obtaining the target video data in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00163] Operation S2026: Recognize text in the target video data.

[00164] Specifically, the server may perform text recognition processing on the target video data of the to-be-recommended media object. For a specific process of performing, by the server, text recognition processing on the target video data, refer to the descriptions of the specific process of recognizing the video text content in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00165] Operation S2027: Output the target speech recognition text.

[00166] Operation S2028: End.

[00167] In this embodiment of the present disclosure, a method for boosting a to-be-processed media object (namely, the to-be-recommended media object) in a cold-start or long-tail case is mainly described. According to the method, when the target audio data of the to-be-recommended media object is obtained, speech recognition processing may be performed on the target audio data to obtain the target speech recognition text. Then, in this embodiment of the

present disclosure, semantic understanding processing may be further performed on the target speech recognition text, and then text vector representation processing is performed on the target speech recognition text to obtain a target text representation vector. Further, in this embodiment of the present disclosure, when N historical text representation vectors for representing N historical media objects are obtained, vector similarity calculation may be separately performed on the target text representation vector and the N historical text representation vectors. Then K historical media objects with highest similarities may be selected based on vector similarities obtained through calculation, to establish a similarity association relationship with the selected K historical media objects. Further, in this embodiment of the present disclosure, K historical recommendation features of the K historical media objects may be obtained based on the similarity association relationship, and then pooling processing may be performed on the K historical recommendation features to obtain a target recommendation feature for representing the to-be-recommended media object. In this way, when performing media recommendation based on the target recommendation feature, the server can enhance boosted recommendation for the to-be-recommended media object.

[00168] For ease of understanding, FIG. 13 is a schematic flowchart of estimating a target service object for a media object according to an embodiment of the present disclosure. In this embodiment of the present disclosure, when an obtained to-be-boosted media object (namely, the foregoing to-be-recommended media object) is a long-tail/cold-start media object 13a, target audio data with capture duration being window duration may be captured in a target sampling window, and then speech recognition processing is performed on the target audio data through an audio data speech recognition module 13b to obtain target speech recognition text. Sampling performed on multimedia data of the to-be-recommended media object by using the target sampling window is essentially real-time dynamic sampling. For a specific process of the real-time dynamic sampling, refer to the descriptions of the specific process of obtaining different audio data through sampling based on sampling windows corresponding to different dynamic sampling time points in the embodiment corresponding to FIG. 3. Details are not described herein again. Further, as shown in FIG. 13, in this embodiment of the present disclosure, a semantic understanding model may perform semantic information extraction on the target speech recognition text based on guiding prompt template data of the audio data speech recognition module 13b, to obtain target semantic understanding information. Then text vector representation processing is performed on the target semantic understanding information to obtain a target text representation vector. A function of the audio data speech recognition module 13b herein is equivalent to that of the foregoing speech conversion understanding component.

[00169] Further, in this embodiment of the present disclosure, a vector similarity calculation

module 13c may obtain N historical text representation vectors for representing N pieces of historical media data, and separately perform vector similarity calculation on the target text representation vector and the N historical text representation vectors. A manner of obtaining the N historical text representation vectors herein is the same as the foregoing manner of obtaining the target text representation vector. However, historical audio data is audio data whose sampling duration is window duration and that is captured based on a historical sampling window (for example, a sampling window corresponding to a case that a historical dynamic sampling time point is an end time point of multimedia data of a historical media object). Similarly, a function of the vector similarity calculation module 13c herein is equivalent to that of the foregoing similarity search component.

[00170] Then, in this embodiment of the present disclosure, K historical media objects with highest similarities may be selected based on N vector similarities obtained through calculation, and a similarity association relationship between the to-be-recommended media object and the K historical media objects is established, to obtain a high-interaction semantic top K list (that is, a similar media object list, the similar media object list being configured for storing K historical media objects that have large interaction data amounts and that have high semantic similarities to the to-be-recommended media object). Further, in this embodiment of the present disclosure, K historical recommendation features of the K historical media objects in the similar media object list may be obtained based on a target recommendation feature determining module 13d, and pooling processing is performed on the K historical recommendation features to obtain a target recommendation feature for boosted recommendation of the to-be-recommended media object. Then, in this embodiment of the present disclosure, a target service object matching the to-be-recommended media object may be found, based on the target recommendation feature by using a recommendation model 13e, from all online users currently accessing a service processing platform (for example, a social software platform), and the to-be-recommended media object is pushed to the currently found target service object. The recommendation model 13e is obtained by training an initial recommendation model by using N pieces of historical interaction behavior information 13f of the N historical media objects.

[00171] Further, FIG. 14 is a schematic flowchart of estimating a target service object for a livestreaming room according to an embodiment of the present disclosure. As shown in FIG. 14, for example, a to-be-recommended media object is a to-be-recommended livestreaming media object (for example, a to-be-boosted livestreaming room). The to-be-boosted livestreaming room may be collectively referred to as a current livestreaming room. In this case, a historical media object may be a historical livestreaming media object (for example, a historical livestreaming room shown in

FIG. 14). The recommendation method for a media object may include at least the following operation S301 to operation S309:

[00172] Operation S301: Time window–based sampling

[00173] Specifically, when multimedia data of a to-be-recommended livestreaming media object is to-be-recommended livestreaming data, a server may first sample the to-be-recommended livestreaming data by using a target sampling window (that is, perform time window–based sampling) to obtain target audio data. For a specific process of obtaining the target audio data by the server through sampling, refer to the descriptions of the specific process of obtaining the target audio data through sampling in the embodiment corresponding to FIG. 3. Details are not described herein again. After operation S301 is performed, operation S303 is performed.

[00174] Operation S302: High-interaction filtering

[00175] Similarly, when multimedia data of a historical media object is historical livestreaming data, the server may select historical media objects with interaction data amounts ranked top 10% (that is, select a high-interaction livestreaming room obtained through high-interaction filtering) based on interaction data curve distribution corresponding to a historical media object database. In other words, in this embodiment of the present disclosure, the interaction data curve distribution may be generated when data volume statistics are collected on interaction data amounts of all historical media objects in the historical media object database, and then top historical livestreaming rooms may be filtered out from the interaction data curve distribution and collectively referred to as high-interaction livestreaming rooms with large interaction data amounts. Then the server may obtain N pieces of historical audio data of N historical livestreaming media objects that are selected. The historical audio data is audio data whose sampling duration is window duration and that is captured by the server based on a historical sampling window (for example, a recorded sampling window in which a historical dynamic sampling time point is an end time point of historical livestreaming data of a historical livestreaming media object). The interaction data curve distribution is a distribution curve chart determined by using each historical media object in the historical media object database and an interaction data amount of the historical media object as two coordinate axes. For details, refer to the descriptions of the specific process of selecting the historical media objects in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00176] Operation S303: Speech recognition processing

[00177] Specifically, in this embodiment of the present disclosure, ASR processing (namely,

speech recognition processing) may be separately performed on the target audio data of the to-be-recommended livestreaming media object and the N pieces of historical audio data of the N historical livestreaming media objects to obtain target speech recognition text and N pieces of historical speech recognition text. For a specific process of performing, by the server, speech recognition processing on historical audio data of each historical media object, refer to the descriptions of the specific process of performing speech recognition processing on the target audio data in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00178] Operation S304: Semantic understanding information extraction

[00179] Specifically, the server may separately perform, through the semantic understanding model (for example, an LLM) based on guiding prompt template data, semantic understanding information extraction (namely, semantic information extraction) on the target speech recognition text and the N pieces of historical speech recognition text to extract target semantic understanding information and N pieces of historical semantic understanding information. The target semantic understanding information is determined based on target semantic extraction information obtained by performing semantic information extraction on the target speech recognition text. Similarly, the historical semantic understanding information is determined based on historical semantic extraction information obtained by performing semantic information extraction on the historical speech recognition text. For a specific process of performing, by the server, semantic information extraction on each piece of historical speech recognition text, refer to the descriptions of the specific process of performing semantic information extraction on the target speech recognition text in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00180] Operation S305: Text vector representation

[00181] Specifically, in this embodiment of the present disclosure, text vector representation processing may be separately performed on the target semantic understanding information and the N pieces of historical semantic understanding information to obtain a target text representation vector corresponding to the target semantic understanding information and a historical text representation vector corresponding to each piece of historical semantic understanding information. For a specific process of performing, by the server, text vector representation processing on each piece of historical semantic understanding information, refer to the descriptions of the specific process of performing text vector representation processing on the target semantic understanding information in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00182] Operation S306: Current livestreaming room representation

[00183] Specifically, in this embodiment of the present disclosure, after text vector representation processing is performed on the target semantic understanding information, text representation vectors obtained through processing may be collectively referred to as the current livestreaming room representation. To be specific, the current livestreaming room representation may be configured for representing a target text representation vector of a current to-be-recommended livestreaming media object (namely, a current livestreaming room).

[00184] Operation S307: High-interaction livestreaming room representation

[00185] Similarly, specifically, in this embodiment of the present disclosure, after text vector representation processing is separately performed on the N pieces of historical semantic understanding information in operation S305, each text representation vector obtained through processing may also be collectively referred to as the high-interaction livestreaming room representation. To be specific, the high-interaction livestreaming room representation is configured for representing a historical text representation vector of each historical livestreaming room that has a large interaction data amount and that is obtained through high-interaction filtering. Further, the server may separately perform vector similarity calculation on the target text representation vector and the N historical text representation vectors, and select K historical text representation vectors with highest similarities from the N historical text representation vectors based on vector similarities obtained through calculation.

[00186] Operation S308: Obtain a livestreaming room sequence.

[00187] Specifically, the server establishes a similarity association relationship between each of K historical livestreaming media objects corresponding to the selected K historical text representation vectors and the to-be-recommended livestreaming media object, to obtain a similar media object list including the K historical livestreaming media objects (the similar media object list herein may be the livestreaming room sequence shown in FIG. 14).

[00188] Operation S309: Recommendation model

[00189] Specifically, in this embodiment of the present disclosure, K historical recommendation features of K historical media objects in the similar media object list may be obtained, and then pooling processing is performed on the K historical recommendation features to obtain a target recommendation feature for representing the to-be-recommended livestreaming media object. Then the server may invoke the recommendation model shown in FIG. 14 to find, based on the target recommendation feature, a target service object matching the to-be-recommended livestreaming media object from all online users (namely, all service objects)

accessing a service processing platform (for example, a social software platform), to recommend the to-be-recommended livestreaming media object (namely, the current livestreaming room shown in FIG. 14) to the target service object through boosted recommendation.

[00190] In this embodiment of the present disclosure, when the target audio data is captured based on the target sampling window, the target text representation vector for representing the to-be-recommended media object may be obtained through calculation based on the target speech recognition text associated with the target audio data. In this way, similarity association can be established between the target text representation vector and historical text representation features of the historical media objects through vector similarity calculation and search, so that the target recommendation feature of the to-be-recommended media object can be obtained through equivalent calculation based on the historical recommendation features of the selected K historical media objects associated with the to-be-recommended media object. In this way, when the to-be-recommended media object has a small interaction data amount, the target recommendation feature of the to-be-recommended media object can still be equivalently described by using the historical recommendation features of the K historical media objects with high semantic similarities. In this way, when performing media recommendation by using a target recommendation model, a recommendation system on the server can directly recommend, based on the target recommendation feature currently obtained through equivalent calculation, the to-be-recommended media object to a target service object matching the to-be-recommended media object, to improve accuracy of media recommendation.

[00191] For ease of understanding, FIG. 15 is a schematic flowchart of a recommendation method for a media object according to an embodiment of the present disclosure. In an embodiment, the recommendation method for a media object may include at least the following operation S401 to operation S407:

[00192] Operation S401: Obtain a target dynamic sampling time point and window duration of a to-be-recommended media object; determine, based on the target dynamic sampling time point and the window duration, a target sampling window for sampling multimedia data of the to-be-recommended media object; and then obtain, from the multimedia data of the to-be-recommended media object based on the target sampling window, audio data with a data sampling end time point being the target dynamic sampling time point and with sampling duration being the window duration, and use the obtained audio data as target audio data of the to-be-recommended media object in the target sampling window.

[00193] The target dynamic sampling time point is a time point changing with a change in

media data association duration of a to-be-processed media object (that is, a to-be-recommended media object that needs boosted recommendation). The target dynamic sampling time point includes a first dynamic sampling time point. The target sampling window includes a first sampling window corresponding to the first dynamic sampling time point. The target audio data includes first audio data captured based on the first dynamic sampling time point and the window duration.

[00194] For ease of understanding, when the to-be-recommended media object is a livestreaming media object and the multimedia data of the to-be-recommended media object is to-be-recommended livestreaming data, a historical media object may be a historical livestreaming media object, and multimedia data of the historical media object may be historical livestreaming data. In this case, when receiving the multimedia data, transmitted by a first local device, of the to-be-recommended media object, a server may first obtain the target dynamic sampling time point and the window duration. The directly obtained target dynamic sampling time point and window duration may be default values uniformly set by the server.

[00195] Further, in this embodiment of the present disclosure, a sampling window for sampling the to-be-recommended livestreaming data may be used as the target sampling window. Then the audio data whose sampling duration is the window duration and that is captured based on the target sampling window is obtained by using the target dynamic sampling time point as an end time point of data sampling, and the audio data obtained in the target sampling window is used as the target audio data.

[00196] The 1st time point used as the target dynamic sampling time point is the first dynamic sampling time point determined by starting from start duration of the to-be-recommended livestreaming data until sampling duration is the window duration, and the first sampling window obtained when the sampling duration is the window duration is the target sampling window. The target sampling window is configured for capturing and collecting the to-be-recommended livestreaming data.

[00197] Operation S402: Obtain sampling interval duration for dynamic sampling; determine, based on the first dynamic sampling time point and the sampling interval duration, a second dynamic sampling time point for sampling the multimedia data of the to-be-recommended media object; then determine, based on the second dynamic sampling time point and the window duration, a second sampling window for sampling the multimedia data of the to-be-recommended media object; then obtain, from the multimedia data of the to-be-recommended media object based on the second sampling window, audio data with a data sampling end time point being the second dynamic sampling time point and with sampling duration being the window duration; and use the obtained

audio data as second audio data of the to-be-recommended media object in the target sampling window, and update the target audio data by using the second audio data when the target dynamic sampling time point is updated by using the second dynamic sampling time point.

[00198] The second dynamic sampling time point is a next sampling time point of the first dynamic sampling time point. The second sampling window is a next sampling window of the first sampling window, and window overlapping exists between the second sampling window and the first sampling window.

[00199] Specifically, the sampling interval duration that is configured for dynamic sampling and that is obtained by the server is interval duration determined based on two adjacent time points that are separately used as the target dynamic sampling time point. In other words, in this embodiment of the present disclosure, the second dynamic sampling time point may alternatively be determined based on the first dynamic sampling time point and the sampling interval duration. That is, interval duration between the second dynamic sampling time point and the first dynamic sampling time point is the sampling interval duration. Further, in this embodiment of the present disclosure, another sampling window (namely, another target sampling window) for sampling the to-be-recommended livestreaming data may be collectively referred to as the second sampling window. Then the target sampling window may be updated by using the second sampling window (to be specific, a second sampling window in which the second dynamic sampling time point is a data sampling end time point and sampling duration is the window duration), to implement data capture of the to-be-recommended livestreaming data. Then the target audio data may be updated by using audio data in livestreaming data captured based on the second sampling window, to achieve dynamic sampling.

[00200] The sampling interval duration is configured for indicating that window overlapping exists between any two adjacent target sampling windows. For ease of understanding, in other words, when the sampling interval duration is less than or equal to window duration of the target sampling window, the to-be-recommended livestreaming data can be completely captured.

[00201] For ease of understanding, FIG. 16 is a diagram of dynamic sampling based on a play timeline of to-be-processed livestreaming data according to an embodiment of the present disclosure. The play timeline of the to-be-processed livestreaming data herein may be configured for indicating a livestreaming progress of a current livestreaming room. As shown in FIG. 16, the target sampling window is specifically a sampling window newly determined from a first sampling window 16b1 and a second sampling window 16b2, and the target dynamic sampling time point may specifically include a first dynamic sampling time point 16a1 and a second dynamic sampling

time point 16a2. For example, as shown in FIG. 16, when a livestreaming progress of the to-be-recommended livestreaming data reaches a moment T2, the moment T2 may be used as the first dynamic sampling time point 16a1, and a sampling window corresponding to the moment T2 may be correspondingly used as the first sampling window 16b1. To be specific, to-be-recommended livestreaming data with a sampling end time point being a time point at the moment T2 and with sampling duration being window duration 16d may be obtained through sampling based on the first sampling window 16b1, and multimedia data ending with a target dynamic sampling time point 16A (namely, a sampling end time point) and with sampling duration being the window duration 16d may be captured based on a target sampling window 16B. Then audio data of the multimedia data captured in the target sampling window 16B is extracted and is used as the target audio data. Further, with proceeding of the to-be-recommended livestreaming data, in this embodiment of the present disclosure, whether interval duration between a current livestreaming progress of the to-be-recommended livestreaming data and the livestreaming progress represented by the moment T2 reaches sampling interval duration 16c may be dynamically determined. If the sampling interval duration 16c is reached, it can be determined, based on the moment T2 and the sampling interval duration 16c, that the livestreaming progress of the to-be-recommended livestreaming data currently reaches a moment T3. Therefore, the moment T3 may be used as the second dynamic sampling time point 16a2, and a sampling window corresponding to the moment T3 may be used as the second sampling window 16b2. Then another piece of to-be-recommended livestreaming data with a sampling end time point being a time point at the moment T3 and with sampling duration being the window duration 16d may be obtained through sampling based on the second sampling window 16b2. Then the second dynamic sampling time point 16a2 may be used as a new target dynamic sampling time point 16A, to replace the sampling end time point at which the first dynamic sampling time point 16a1 is the target dynamic sampling time point 16A. Therefore, the first sampling window 16b1 may be replaced with the second sampling window 16b2 used as a new target sampling window, to dynamically update the target sampling window 16B. Then audio data extracted from the to-be-recommended livestreaming data in the second sampling window 16b2 is used as another piece of target audio data, to perform the following operation S403 to operation 407. In one embodiment, to avoid a waste of storage resources, in this embodiment of the present disclosure, when new audio data (namely, the another piece of target audio data) is captured based on the second sampling window 16b2, old audio data originally captured based on the first sampling window 16b1 is further replaced with the new audio data, that is, the new audio data may be directly used as current latest target audio data.

[00202] Operation S403: Perform text vector representation processing on target speech

recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the to-be-recommended media object.

[00203] A speech conversion understanding component and a similarity search component for performing speech conversion processing on the target audio data are obtained. The speech conversion understanding component includes a speech recognizer and a semantic understanding model for performing semantic extraction, and the similarity search component includes a text representer.

[00204] Further, speech recognition processing is performed on the target audio data through the speech recognizer in the speech conversion understanding component to obtain the target speech recognition text associated with the target audio data. Then information extraction is performed on the target speech recognition text through the semantic understanding model in the speech conversion understanding component to obtain target semantic understanding information. Further, text vector representation processing is performed on the target semantic understanding information through the text representer in the similarity search component to obtain a target text representation vector for representing a to-be-recommended livestreaming media object.

[00205] Operation S404: Obtain N historical text representation vectors for representing N historical media objects, and separately perform vector similarity calculation on the target text representation vector and each historical text representation vector to obtain N vector similarities.

[00206] N is a positive integer greater than 1. One historical text representation vector is recorded after text vector representation processing is performed on historical speech recognition text corresponding to one piece of historical audio data, and one piece of historical audio data is audio data of one historical media object in a recorded historical sampling window. Window duration of the historical sampling window is consistent with window duration of the target sampling window.

[00207] Specifically, the target text representation vector is obtained by performing speech conversion processing on the target audio data through the speech conversion understanding component and the similarity search component for speech conversion processing. Then the N historical text representation vectors for representing the N historical media objects are obtained, and after the N historical text representation vectors are obtained, vector similarity calculation is separately performed on the target text representation vector for representing the to-be-recommended livestreaming media object and N historical text representation vectors for representing N historical livestreaming media objects. A vector similarity calculation method includes a cosine similarity, a Jaccard coefficient similarity, a Euclidean distance, a Manhattan

distance, and the like. This is not limited herein.

[00208] Operation S405: Select, based on the N vector similarities, K historical media objects from the N historical media objects, and establish a similarity association relationship between K historical media objects and the to-be-recommended media object.

[00209] Specifically, the N historical livestreaming media objects is sorted based on the N vector similarities obtained through calculation. Then K historical livestreaming media objects associated with the to-be-recommended livestreaming media object are selected, and a similarity association relationship between the K historical livestreaming media objects and the to-be-recommended livestreaming media object is established. For a specific implementation, refer to the descriptions of the specific process of selecting the K historical media objects in the embodiment corresponding to FIG. 3. Details are not described herein again.

[00210] Operation S406: Obtain K historical recommendation features of the K historical media objects from a recommendation feature database associated with the K historical media objects.

[00211] Specifically, K historical recommendation features of the K historical livestreaming media objects are obtained by the server from a recommendation feature database associated with the K historical livestreaming media objects.

[00212] The recommendation feature database is obtained by using the following method:

[00213] obtaining N pieces of historical interaction behavior information corresponding to the N historical media objects, one historical media object corresponding to one piece of historical interaction behavior information, and one piece of historical interaction behavior information being configured for representing item interaction details of each interaction item in one historical media object; using item interaction details represented by the N pieces of historical interaction behavior information as a training sample for training an initial recommendation model, and obtaining a recommendation indicator for training the initial recommendation model; performing model training on the initial recommendation model by using the training sample and the recommendation indicator to obtain a target recommendation model, the target recommendation model being configured for extracting the N historical recommendation features of the N historical media objects, and one historical recommendation feature being obtained by performing, by the target recommendation model, interaction vector extraction on item interaction details represented by one piece of historical interaction behavior information; and adding the N historical recommendation features extracted by the target recommendation model to the recommendation feature database.

[00214] Specifically, the server obtains historical interaction behavior information corresponding to each of the N historical livestreaming media objects, and then uses user-item interaction details represented by the obtained historical interaction behavior information of the N historical livestreaming media objects as a training sample for an initial recommendation model. The initial recommendation model is an untrained model that is configured for performing boosted recommendation for a long-tail or cold-start media object. During training of the initial recommendation model, a recommendation indicator that is associated with the initial recommendation model and that is configured for training is obtained.

[00215] Further, the initial recommendation model is trained based on the recommendation indicator of the initial recommendation model to obtain a target recommendation model for extracting a recommendation feature of the to-be-recommended media object and a recommendation feature of a historical media object. The N historical livestreaming media objects is processed through the target recommendation model to obtain N historical recommendation features of the N historical livestreaming media objects, and the N historical recommendation features are added to the recommendation feature database.

[00216] The to-be-recommended media object (for example, the to-be-recommended livestreaming media object) is a multimedia object, on a service processing platform, that has an interaction data amount less than an interaction data amount threshold. That is, the to-be-recommended media object (for example, the to-be-recommended livestreaming media object) is a cold-start/long-tail media object.

[00217] Operation S407: When the K historical recommendation features of the K historical media objects are obtained, perform feature pooling processing on the K historical recommendation features based on the similarity association relationship to obtain a pooling processing feature associated with the K historical recommendation features, determine the pooling processing feature as the target recommendation feature for representing the to-be-recommended media object, and perform media recommendation for the to-be-recommended media object based on the target recommendation feature.

[00218] Before pooling processing is performed on the K historical recommendation features of the K historical livestreaming media objects, a pooling processing policy for performing pooling processing may be first obtained based on the similarity association relationship. The pooling processing policy includes an average pooling processing policy, a max pooling processing policy, a random pooling processing policy, and the like. This is not limited herein.

[00219] If the pooling processing policy used in the pooling processing is the average

pooling processing policy, average pooling processing may be performed on K historical recommendation features of K pieces of historical livestreaming data through a recommendation model according to the average pooling processing policy, to obtain an average pooling feature. Then the target recommendation feature of the to-be-recommended livestreaming data is obtained based on the average pooling feature.

[00220] For a specific implementation of performing, by the recommendation model, pooling processing on the K historical recommendation features to obtain the target recommendation feature of the to-be-recommended media object, refer to the following formula (9):

$$\hat{r}_i = \frac{1}{K}\sum_{k=1}^{K} M(J_k) \qquad (9)$$

[00221] The subscript k of the symbol J represents an ID index of a k^{th} historical high-interaction media object in a semantic proximity list (namely, a similar media object list) determined for a current livestreaming room (for example, a livestreaming room i). k is a positive integer less than or equal to K, that is, a value of k is 1 to K. K represents K most similar livestreaming rooms found through vector similarity search. M represents a currently trained recommendation model (namely, the target recommendation model) that is deployed in a recommendation scenario for a to-be-recommended media object. $M(J_k)$ represents a recommendation representation, extracted by the target recommendation model, of the k^{th} historical media object.

[00222] Further, when the target recommendation model is deployed and applied on a service processing platform provided by the server, the server may obtain object association data information of a service object accessing the service processing platform (for example, the server obtains historical behavior interaction information transmitted by a second local device under authorization of each service object, and when a recommendation service is a livestreaming recommendation service, the historical behavior interaction information herein is determined based on an interaction operation performed by a user on some items in one or more livestreaming rooms, the interaction operation herein including, for example, any one of a like operation, a favorite operation, and an item transaction operation). The object association data information and the target recommendation feature are inputted to the target recommendation model. When an object vector representation feature for representing a service object is extracted from the object association data information through the target recommendation model (the object vector representation feature herein may be configured for indicating a preference of the service object for items in some livestreaming rooms), a target object vector representation feature matching the target recommendation feature is selected from the object vector representation feature. In this way, the

server can determine a service object corresponding to the target object vector representation feature as a target service object that is selected from the service object and that matches the to-be-recommended media object, and recommend the to-be-recommended media object to the target service object.

[00223] Specifically, the service processing platform includes a social software platform. When the service processing platform is the social software platform, the trained target recommendation model is deployed on the social software platform. In addition, when the target recommendation model is deployed, the selected target service object may receive, through the second local device (for example, the user terminal 20a in the embodiment corresponding to FIG. 1), a to-be-recommended media object that is recommended by the server (for example, the server 10d in the embodiment corresponding to FIG. 1) through boosted recommendation, and may view media data (for example, the foregoing to-be-recommended livestreaming data) of the to-be-recommended media object on the social software platform of the second local device. For example, the server may obtain historical behavior interaction information of all users (namely, the service object) that currently access the social software platform. The historical behavior interaction information herein may be configured for recording item transaction information and user preferences (namely, the object association data information) of the users. Further, the server inputs the target recommendation feature and the obtained item transaction information and user preferences of the users to the target recommendation model, so that the target recommendation model extracts, from the item transaction information and the user preferences, object vector representation features for representing the users. Then an object vector representation feature matching the target recommendation feature is selected from the object vector representation features of all the users of the social software platform, and is used as the target object vector representation feature. Further, a user corresponding to the target object vector representation feature is used as a target service object matching the to-be-recommended livestreaming data, and then the to-be-recommended media object is recommended to the target service object.

[00224] One embodiment of searching, by the target recommendation model, for the target service object matching the to-be-recommended media object is as follows:

$$y_{u,i} = similarity\left(\hat{r}_i, M(u)\right) \qquad (10)$$

[00225] $M(u)$ represents a representation (that is, an object vector representation feature) of a user (for example, an online user u) in a recommendation model M. Finally, a degree of feature matching between the target recommendation feature $\hat{r}_i$ and the object vector representation feature is calculated in a feature matching manner indicated by the formula (10), to evaluate a

preference score $y_{u,i}$ of the online user u for a cold-start/long-tail livestreaming room, to determine a scale of users who meet a requirement, and further achieve recommendation accuracy of personalized boosted recommendation.

[00226] The object association data information and the target recommendation feature are inputted to the target recommendation model, and information vector representation extraction is performed on the object association data information through the target recommendation model to obtain the object vector representation feature for representing the service object. Feature matching is performed between the object vector representation feature and the target recommendation feature to obtain a feature matching result. If the feature matching result indicates that the object vector representation feature includes an object vector representation feature matching the target recommendation feature, the object vector representation feature matching the target recommendation feature is determined as the target object vector representation feature.

[00227] Specifically, when the item transaction information and the user preferences (namely, the object association data information) of the users (namely, the service object) of the social software platform (namely, the service processing platform) and the target recommendation feature are inputted to the target recommendation model deployed on the social software platform, the target recommendation model performs information vector representation extraction on the item transaction information and the user preferences of the users to extract an object vector representation feature for representing user information. Further, a user associated with the to-be-recommended media object is selected from all the users, and then the target recommendation model performs matching between the object vector representation feature and the target recommendation feature of the to-be-recommended media object to obtain a feature matching result. Correspondingly, if the feature matching result indicates that the object vector representation feature includes a vector representation feature matching the target recommendation feature, the object vector representation feature is used as the target object vector representation feature.

[00228] Further, if the to-be-recommended media object is a livestreaming media object, the to-be-recommended livestreaming media object is used as a target recommendation object to be recommended to the target service object in the livestreaming recommendation service, and the target recommendation object is transmitted to a terminal device corresponding to the target service object, to enable the terminal device to display the target recommendation object in a livestreaming recommendation list corresponding to the livestreaming recommendation service. In some embodiments, if the to-be-recommended media object is a video-on-demand media object, a video-on-demand link corresponding to the video-on-demand media object is packaged into a

recommendation notification message, and the recommendation notification message is transmitted to a terminal device corresponding to the target service object, to enable the terminal device to display, in a livestreaming recommendation list corresponding to the livestreaming recommendation service, the video-on-demand link in the recommendation notification message. The video-on-demand link is configured for indicating the terminal device to display the video-on-demand media object in response to a trigger operation on the recommendation notification message.

[00229] To comprehensively evaluate effectiveness of the method provided in the embodiments of the present disclosure, when the to-be-recommended media object is a livestreaming media object, the solution is sufficiently evaluated in two scenarios of the service processing platform: an offline scenario and an online scenario. The ASR algorithm used in the evaluation can be configured for speech recognition, and the used LLM may be configured for semantic understanding. In addition, a quantity of historical livestreaming media objects that have high semantic similarities and that are selected by using a vector similarity algorithm (for example, a cosine similarity) is K. In this way, when K=20, 20 historical livestreaming media objects associated with the to-be-recommended livestreaming media object are selected.

[00230] First, during offline evaluation, attention is mainly paid to three indicators of the target recommendation model: an area under a receiver operating characteristic (ROC) curve (AUC), an area enclosed by a coordinate axis and each group of ROC curves after weighted averaging (group AUC, GAUC), and a precision at cut-off (PCOC).

[00231] In the area under the ROC curve, the ROC curve is a curve drawn by using a false positive rate (FPR) as a horizontal axis and a true positive rate (TPR) as a vertical axis. An AUC value ranges from 0 to 1. A larger AUC value indicates a stronger capability of the model in distinguishing between positive and negative samples. For example, in the embodiments of the present disclosure, an optimal division threshold for performing binary classification (for example, a head livestreaming room with a large interaction data amount or a tail livestreaming room with a small interaction data amount) can be intelligently determined based on the ROC curve, and then the determined optimal division threshold may be used as the foregoing interaction data amount threshold configured in the recommendation system. In other words, in the embodiments of the present disclosure, the AUC value may be calculated based on the area under the ROC curve, and then the AUC value may indicate a capability of the target recommendation model to perform click/tap prediction for a positive sample (for example, a low-interaction livestreaming room or a low-popularity livestreaming room such as a cold-start livestreaming room and/or a long-tail livestreaming room) and a negative sample (for example, a high-interaction livestreaming room or a

high-popularity livestreaming room) when the target recommendation model is obtained through offline training by using training samples of different service objects (for example, historical behavior interaction information of users). For details, refer to Table 1 below. The positive and negative samples herein may be determined according to an actual training requirement. For example, the positive sample may be a high-popularity livestreaming room, and the negative sample may be a cold-start livestreaming room. Specific types of the positive and negative samples participating in the training to obtain the target recommendation model are not limited herein.

[00232] The GAUC is an extension of the AUC, and is usually configured for evaluating performance of a recommendation system on different service object groups or different categories. The GAUC is to perform weighted averaging on each group of AUCs by considering sample sizes of different service object groups. A larger GAUC is better, and indicates balanced performance of the model on the groups. For example, as shown in Table 1 below, livestreaming rooms with click-through page views (PVs) (that is, click-through volumes or interaction data amounts) ranging from 0 to 100 may be divided into one group, and livestreaming rooms with click-through PVs ranging from 100 to 500 may be divided into another group.

[00233] The PCOC is accuracy of the first X to-be-recommended media objects in a recommendation list that have been recommended. The PCOC measures a proportion of media objects, among the first X recommended media objects, in which service objects are actually interested. A larger PCOC is better, and indicates higher accuracy of the recommendation system in terms of the first X to-be-recommended media objects that have been recommended.

[00234] In the following table, the method provided in the embodiments of the present disclosure is evaluated based on PVs in the click-through PVs. In the embodiments of the present disclosure, for relative gains obtained for livestreaming rooms of different popularity with respect to a baseline after the method provided in the embodiments of the present disclosure is used, refer to Table 1 below. The livestreaming room is a to-be-recommended livestreaming media object.

[00235] Table 1

Click-through PVs	**0 to 100**	**100 to 500**	**500 to 2000**	**2000 to 5000**	**5000+**
AUC	+0.049%	+0.047%	+0.023%	+0.015%	+0.014%
GAUC	+0.149%	+0.079%	+0.034%	+0.019%	+0.016%
PCOC	+0.189%	+0.112%	+0.064%	+0.043%	+0.042%

[00236] As shown in Table 1, when the embodiments of the present disclosure are implemented in livestreaming rooms of different popularity, different gains are obtained with

respect to the baseline (to be specific, a click-through volume predicted when these indicators are evaluated by using a recommendation model in the related art).

[00237] Further, in the embodiments of the present disclosure, during online evaluation, attention is mainly paid to boost effect for a livestreaming room in a cold-start channel or a long-tail channel (that is, the foregoing to-be-recommended livestreaming media object). In the embodiments of the present disclosure, for relative benefits of cold-start channel distribution with respect to a baseline, refer to Table 2 below.

[00238] Table 2

Cold-start distribution	**PVs**	**PVs in 1 min**	**PV ratio in 1 min**
Proximity search=10	+0.866%	+0.629%	+0.825%
Proximity search=20	+1.524%	+1.173%	+1.248%

[00239] In the embodiments of the present disclosure, for relative benefits of long-tail content distribution with respect to a baseline, refer to Table 3 below.

[00240] Table 3

Long-tail content distribution	**PVs**	**PVs in 1 min**	**PV ratio in 1 min**
PVs < 500 views	+2.698%	+6.153%	+3.287%
PVs < 100 views	+4.034%	+6.721%	+2.658%

[00241] As shown in Table 2 and Table 3, in the embodiments of the present disclosure, different gains are obtained in the cold-start channel distribution and the long-tail content distribution with respect to the baseline.

[00242] In the embodiments of the present disclosure, audio data, with a data sampling end time point being a target dynamic sampling time point and with sampling duration being window duration, that is captured from multimedia data of a to-be-recommended media object based on a target sampling window is used as target audio data. In addition, the target dynamic sampling time point is updated for the target audio data based on sampling interval duration, and then the target audio data associated with the to-be-recommended media object is updated. Further, the target audio data obtained in the target sampling window may be recognized to obtain target speech recognition text associated with the target audio data. Further, in the embodiments of the present disclosure, text vector representation processing is performed on the target speech recognition text associated with the audio data to obtain a target text representation vector for representing the to-be-recommended media object. N historical text representation vectors for representing N historical media objects are

obtained, and then vector similarity calculation is separately performed on the target text representation vector and each of the N historical text representation vectors to obtain N vector similarities. Further, in the embodiments of the present disclosure, K historical media objects highly similar to the to-be-recommended media object are selected from the N historical media objects based on the N vector similarities, and then a similarity association relationship between the to-be-recommended media object and the selected K historical media objects is established. In the embodiments of the present disclosure, K historical recommendation features of the K historical media objects are obtained from a recommendation feature database associated with the K historical media objects, and then a pooling processing policy for performing pooling processing is obtained to perform pooling processing on the K historical recommendation features to obtain a target recommendation feature of the to-be-recommended media object. Further, historical interaction behavior information of the N historical media objects is obtained. Therefore, when a target recommendation model is deployed and applied on a service processing platform, object association data information of a service object accessing the service processing platform is obtained, then a target service object matching the to-be-recommended media object is selected from the service object of the service processing platform based on the object association data information and the target recommendation feature, and then media recommendation is performed for the target service object. In the embodiments of the present disclosure, the multimedia data of the to-be-recommended media object may be dynamically sampled to obtain the target audio data associated with the to-be-recommended media object, content in the target audio data is vectorized, and similarity calculation is performed with a vector of content in historical audio data, to find the K historical media objects highly associated with the to-be-recommended media object, so that the historical recommendation features of the K historical media objects are obtained from the recommendation feature database associated with the K historical media objects. Then pooling processing is performed on the historical recommendation features to obtain the target recommendation feature for representing the to-be-recommended media object. Then the target service object matching the to-be-recommended media object is selected based on N pieces of historical interaction behavior information of the N historical media objects and the target recommendation feature of the to-be-recommended media object, and to-be-recommended media data is pushed to the target service object. In this way, accuracy of performing, by the target recommendation model, matching between a cold-start to-be-recommended media object and a service object is improved, and the to-be-recommended media object is further boosted.

[00243] Further, FIG. 17 is a diagram of a structure of a recommendation apparatus for a media object according to an embodiment of the present disclosure. The recommendation apparatus

for a media object may be a computer program (including program code) running on a computer device. For example, the recommendation apparatus 1 for a media object may be application software, and the recommendation apparatus 1 for a media object may be configured to perform corresponding operations in the method provided in the embodiments of the present disclosure. The recommendation apparatus 1 for a media object may include an audio data speech recognition module 11, a vector similarity calculation module 12, a similarity association relationship establishment module 13, and a target recommendation feature determining module 14.

[00244] The audio data speech recognition module 11 is configured to obtain target audio data of a to-be-recommended media object in a target sampling window, and perform text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the to-be-recommended media object.

[00245] The vector similarity calculation module 12 is configured to obtain N historical text representation vectors for representing N historical media objects, and separately perform vector similarity calculation on the target text representation vector and the N historical text representation vectors to obtain N vector similarities, an interaction data amount of each historical media object being greater than an interaction data amount of the to-be-recommended media object, N being a positive integer greater than 1, each of the N historical text representation vectors being recorded after text vector representation processing is performed on historical speech recognition text corresponding to each piece of historical audio data, each piece of historical audio data being audio data of each historical media object in a recorded historical sampling window, and window duration of the historical sampling window being consistent with window duration of the target sampling window.

[00246] The similarity association relationship establishment module 13 is configured to establish a similarity association relationship between K historical media objects and the to-be-recommended media object when the K historical media objects are selected from the N historical media objects based on the N vector similarities, K being a positive integer less than N.

[00247] The target recommendation feature determining module 14 is configured to determine, based on the similarity association relationship and K historical recommendation features when the K historical recommendation features of the K historical media objects are obtained, a target recommendation feature for representing the to-be-recommended media object, and perform media recommendation for the to-be-recommended media object based on the target recommendation feature.

[00248] For exemplary implementations of the audio data speech recognition module 11, the

vector similarity calculation module 12, the similarity association relationship establishment module 13, and the target recommendation feature determining module 14, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00249] The audio data speech recognition module 11 includes a target audio data determining unit 110, a target audio data update unit 111, a component obtaining unit 112, a speech recognition unit 113, a semantic understanding unit 114, a text vector representation unit 115, an invalid audio data marking unit 116, and a target speech recognition text update unit 117.

[00250] The target audio data determining unit 110 is configured to obtain a target dynamic sampling time point and window duration of the to-be-recommended media object, and determine, based on the target dynamic sampling time point and the window duration, the target sampling window for sampling multimedia data of the to-be-recommended media object, the target dynamic sampling time point being a time point changing with a change in media data association duration of the to-be-recommended media object.

[00251] The target audio data determining unit 110 is further configured to obtain, from the multimedia data of the to-be-recommended media object based on the target sampling window, audio data with a data sampling end time point being the target dynamic sampling time point and with sampling duration being the window duration.

[00252] The target audio data determining unit 110 is further configured to use the obtained audio data as the target audio data of the to-be-recommended media object in the target sampling window.

[00253] In some embodiments, the target dynamic sampling time point includes a first dynamic sampling time point, the target sampling window includes a first sampling window corresponding to the first dynamic sampling time point, and the target audio data includes first audio data captured based on the first dynamic sampling time point and the window duration;

[00254] the target audio data update unit 111 is configured to obtain sampling interval duration for dynamic sampling, and determine, based on the first dynamic sampling time point and the sampling interval duration, a second dynamic sampling time point for sampling the multimedia data of the to-be-recommended media object, the second dynamic sampling time point being a next sampling time point of the first dynamic sampling time point;

[00255] the target audio data update unit 111 is further configured to determine, based on the

second dynamic sampling time point and the window duration, a second sampling window for sampling the multimedia data of the to-be-recommended media object, the second sampling window being a next sampling window of the first sampling window, and window overlapping exists between the second sampling window and the first sampling window;

[00256] the target audio data update unit 111 is further configured to obtain, from the multimedia data of the to-be-recommended media object based on the second sampling window, audio data with a data sampling end time point being the second dynamic sampling time point and with sampling duration being window duration; and

[00257] the target audio data update unit 111 is further configured to use the obtained audio data as second audio data of the to-be-recommended media object in the target sampling window, and update the target audio data by using the second audio data when the target dynamic sampling time point is updated by using the second dynamic sampling time point.

[00258] Specifically, the component obtaining unit 112 is configured to obtain a speech conversion understanding component and a similarity search component for performing speech conversion processing on the target audio data, the speech conversion understanding component including a speech recognizer and a semantic understanding model for performing semantic extraction, and the similarity search component including a text representer;

[00259] the speech recognition unit 113 is configured to perform speech recognition processing on the target audio data through the speech recognizer to obtain a speech recognition processing result;

[00260] the semantic understanding unit 114 is configured to: if the speech recognition processing result indicates that the target speech recognition text of the target audio data is recognized, perform text semantic understanding processing on the target speech recognition text through the semantic understanding model to obtain target semantic understanding information of the target speech recognition text; and

[00261] the text vector representation unit 115 is configured to perform vector representation on the target semantic understanding information through the text representer to obtain the target text representation vector for representing the to-be-recommended media object.

[00262] In some embodiments, the invalid audio data marking unit 116 is configured to: if the speech recognition processing result indicates that no target speech recognition text of the target audio data is recognized, mark the target audio data as invalid audio data.

[00263] In some embodiments, the target speech recognition text update unit 117 is

configured to obtain video data of the to-be-recommended media object from the multimedia data of the to-be-recommended media object based on the target sampling window, and use the obtained video data as target video data of the to-be-recommended media object in the target sampling window; and

[00264] the target speech recognition text update unit 117 is further configured to perform video content recognition on a video frame in the target video data to obtain video text content for representing the target video data, and update the target speech recognition text by using the video text content.

[00265] Specifically, the semantic understanding unit 114 is specifically configured to obtain guiding prompt template data for guiding the semantic understanding model to generate a model output, input the guiding prompt template data and the target speech recognition text to the semantic understanding model, perform semantic information extraction on the target speech recognition text through the semantic understanding model based on data content in the guiding prompt template data to obtain target semantic extraction information of the target speech recognition text, and determine the target semantic understanding information of the target speech recognition text based on the target semantic extraction information.

[00266] Further, the semantic understanding unit 114 is specifically configured to: if an object type of the to-be-recommended media object is a video livestreaming type, use prompt template data associated with the video livestreaming type as the guiding prompt template data; and

[00267] the semantic understanding unit 114 is further specifically configured to: if an object type of the to-be-recommended media object is a video-on-demand type, use prompt template data associated with the video-on-demand type as the guiding prompt template data.

[00268] For exemplary implementations of the target audio data determining unit 110, the target audio data update unit 111, the component obtaining unit 112, the speech recognition unit 113, the semantic understanding unit 114, the text vector representation unit 115, the invalid audio data marking unit 116, and the target speech recognition text update unit 117, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00269] Further, the vector similarity calculation module 12 includes an interaction data curve distribution obtaining unit 121, a historical media object determining unit 122, and a historical text representation vector obtaining unit 123.

[00270] The interaction data curve distribution obtaining unit 121 is configured to obtain interaction data curve distribution corresponding to a historical media object database, the interaction data curve distribution being determined based on each historical media object in the historical media object database and an interaction data amount of the historical media object.

[00271] The historical media object determining unit 122 is configured to obtain, from the historical media object database based on the interaction data curve distribution, the N historical media objects with interaction data amounts reaching an interaction data threshold.

[00272] The historical text representation vector obtaining unit 123 is configured to obtain the N historical text representation vectors of the N historical media objects from a historical text vector representation database associated with the N historical media objects, each historical text representation vector being obtained by performing speech conversion processing on historical audio data of each historical media object through the speech conversion understanding component and the similarity search component.

[00273] For exemplary implementations of the interaction data curve distribution obtaining unit 121, the historical media object determining unit 122, and the historical text representation vector obtaining unit 123, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00274] The similarity association relationship establishment module 13 includes a historical media object list determining unit 131 and a similarity association relationship establishment unit 132.

[00275] The historical media object list determining unit 131 is configured to sort the N historical media objects based on the N vector similarities to obtain a historical media object list associated with the N historical media objects.

[00276] The similarity association relationship establishment unit 132 is configured to obtain, from the historical media object list, the K historical media objects with vector similarities meeting a similarity retrieval condition, and establish the similarity association relationship between the K historical media objects and the to-be-recommended media object, the similarity association relationship being configured for representing that a list formed by the K historical media objects is a similar media object list, and any media object in the similar media object list has an association relationship with the to-be-recommended media object.

[00277] For exemplary implementations of the historical media object list determining unit

131 and the similarity association relationship establishment unit 132, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00278] The target recommendation feature determining module 14 includes a target recommendation feature determining unit 141 and a pooling processing unit 142.

[00279] The target recommendation feature determining unit 141 is configured to: when the K historical recommendation features of the K historical media objects are obtained, perform feature pooling processing on the K historical recommendation features based on the similarity association relationship to obtain a pooling processing feature associated with the K historical recommendation features.

[00280] The pooling processing unit 142 is configured to determine the pooling processing feature as the target recommendation feature for representing the to-be-recommended media object.

[00281] The pooling processing unit 142 is specifically configured to obtain, based on the similarity association relationship, a pooling processing policy for performing pooling processing.

[00282] The pooling processing unit 142 is further specifically configured to: when the pooling processing policy is an average pooling processing policy, perform average pooling processing on the K historical recommendation features to obtain an average pooling feature, the average pooling feature being a pooling processing feature associated with the K historical recommendation features.

[00283] For exemplary implementations of the target recommendation feature determining unit 141 and the pooling processing unit 142, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00284] The apparatus 1 further includes a historical recommendation feature obtaining module 15.

[00285] The historical recommendation feature obtaining module 15 is configured to obtain the K historical recommendation features of the K historical media objects from a recommendation feature database associated with the K historical media objects.

[00286] The historical recommendation feature obtaining module 15 includes a historical interaction behavior information obtaining unit 151, a recommendation indicator obtaining unit 152, a target recommendation model training unit 153, and a recommendation feature database update unit 154.

[00287] The historical interaction behavior information obtaining unit 151 is configured to obtain N pieces of historical interaction behavior information corresponding to the N historical media objects, one historical media object corresponding to one piece of historical interaction behavior information, and one piece of historical interaction behavior information being configured for representing item interaction details of each interaction item in one historical media object.

[00288] The recommendation indicator obtaining unit 152 is configured to use item interaction details represented by the N pieces of historical interaction behavior information as a training sample for training an initial recommendation model, and obtain a recommendation indicator for training the initial recommendation model.

[00289] The target recommendation model training unit 153 is configured to perform model training on the initial recommendation model by using the training sample and the recommendation indicator to obtain a target recommendation model, the target recommendation model being configured for extracting the N historical recommendation features of the N historical media objects, and one historical recommendation feature being obtained by performing, by the target recommendation model, interaction vector extraction on item interaction details represented by one piece of historical interaction behavior information.

[00290] The recommendation feature database update unit 154 is configured to add the N historical recommendation features extracted by the target recommendation model to the recommendation feature database.

[00291] For exemplary implementations of the historical interaction behavior information obtaining unit 151, the recommendation indicator obtaining unit 152, the target recommendation model training unit 153, and the recommendation feature database update unit 154, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again.

[00292] Further, the to-be-recommended media object is a multimedia object, on a service processing platform, that has an interaction data amount less than an interaction data amount threshold, a recommendation model corresponding to the service processing platform is the target recommendation model, and the target recommendation model is obtained by performing, based on the N pieces of historical interaction behavior information corresponding to the N historical media objects, model training on an initial recommendation model deployed on the service processing platform.

[00293] The target recommendation feature determining module 14 includes an association data information determining unit 143, an object vector feature determining unit 144, and a media object recommendation unit 145.

[00294] The association data information determining unit 143 is configured to: when the target recommendation model is deployed and applied on the service processing platform, obtain object association data information of a service object accessing the service processing platform.

[00295] The object vector feature determining unit 144 is configured to input the object association data information and the target recommendation feature to the target recommendation model, and when an object vector representation feature for representing the service object is extracted from the object association data information through the target recommendation model, select, from the object vector representation feature, a target object vector representation feature matching the target recommendation feature.

[00296] The media object recommendation unit 145 is configured to determine a service object corresponding to the target object vector representation feature as a target service object that is selected from the service object and that matches the to-be-recommended media object, and recommend the to-be-recommended media object to the target service object.

[00297] Specifically, the object vector feature determining unit 144 is specifically configured to input the object association data information and the target recommendation feature to the target recommendation model, and perform information vector representation extraction on the object association data information through the target recommendation model to obtain the object vector representation feature for representing the service object.

[00298] The object vector feature determining unit 144 is further specifically configured to perform feature matching between the object vector representation feature and the target recommendation feature to obtain a feature matching result.

[00299] The object vector feature determining unit 144 is further specifically configured to: if the feature matching result indicates that the object vector representation feature includes an object vector representation feature matching the target recommendation feature, determine the object vector representation feature matching the target recommendation feature as the target object vector representation feature.

[00300] Specifically, the media object recommendation unit 145 is specifically configured to: if the to-be-recommended media object is a livestreaming media object, use the livestreaming media object as a target recommendation object to be recommended to the target service object in the

livestreaming recommendation service, and transmit the target recommendation object to a terminal device corresponding to the target service object, to enable the terminal device to display the target recommendation object in a livestreaming recommendation list corresponding to the livestreaming recommendation service.

[00301] The media object recommendation unit 145 is further specifically configured to: if the to-be-recommended media object is a video-on-demand media object, package a video-on-demand link corresponding to the video-on-demand media object into a recommendation notification message, and transmit the recommendation notification message to a terminal device corresponding to the target service object, to enable the terminal device to display, in a livestreaming recommendation list corresponding to the livestreaming recommendation service, the video-on-demand link in the recommendation notification message, the video-on-demand link being configured for indicating the terminal device to display the video-on-demand media object in response to a trigger operation on the recommendation notification message.

[00302] For exemplary implementations of the association data information determining unit 143, the object vector feature determining unit 144, and the media object recommendation unit 145, refer to the descriptions of operation S101 to operation S104 in the embodiment corresponding to FIG. 3 and operation S401 to operation S407 in the embodiment corresponding to FIG. 15. Details are not described herein again. In addition, beneficial effects of the same method are not described herein again.

[00303] In exemplary implementations of the present disclosure, service data of objects such as users, enterprises, or institutions (for example, historical behavior interaction information of the users, object association data information, item transaction information of the users, or user preferences) may be involved. When the foregoing embodiments of the present disclosure are applied to a specific product or technology, permission or consent from the objects such as the users, the enterprises, or the institutions needs to be obtained, and collection, use, and processing of related data need to comply with related laws, regulations, and standards in related countries and regions.

[00304] Further, FIG. 18 is a diagram of a structure of a computer device according to an embodiment of the present disclosure. As shown in FIG. 18, the computer device 1000 may be a user terminal, for example, the user terminal 10a or the user terminal 20a in the embodiment corresponding to FIG. 1, or may be a server, for example, the server 10d in the embodiment corresponding to FIG. 1. This is not limited herein. For ease of understanding, an example in which the computer device is a user terminal is used in the present disclosure. The computer device 1000

may include a processor 1001, a network interface 1004, and a memory 1005. In addition, the computer device 1000 may further include a user interface 1003 and at least one communication bus 1002. The communication bus 1002 is configured to implement connection and communication between these components. The user interface 1003 may further include a standard wired interface and a standard wireless interface. In some embodiments, the network interface 1004 may include a standard wired interface and a standard wireless interface (for example, a wireless fidelity (Wi-Fi) interface). The memory 1005 may be a high-speed random access memory (RAM), or may be a non-volatile memory, for example, at least one magnetic disk memory. In some embodiments, the memory 1005 may alternatively be at least one storage apparatus located away from the processor 1001. As shown in FIG. 18, the memory 1005 used as a computer-readable storage medium may include an operating system, a network communication module, a user interface module, and a device-control application program.

[00305] The network interface 1004 in the computer device 1000 may further provide a network communication function. In some embodiments, the user interface 1003 may further include a display and a keyboard. In the computer device 1000 shown in FIG. 18, the network interface 1004 may provide a network communication function. The user interface 1003 is mainly configured to provide an input interface for a user. The processor 1001 may be configured to invoke the device-control application program stored in the memory 1005, to boost a media object in a cold-start or long-tail case.

[00306] The computer device 1000 described in this embodiment of the present disclosure may implement the descriptions of the recommendation method for a media object in the embodiment corresponding to FIG. 3 or FIG. 15, or may implement the descriptions of the recommendation apparatus 1 in the embodiment corresponding to FIG. 17. Details are not described herein again. In addition, beneficial effects of the same method are not described herein again.

[00307] In addition, the embodiments of the present disclosure further provide a computer-readable storage medium. The computer-readable storage medium stores a computer program to be executed by the foregoing recommendation apparatus 1 for a media object, and the computer program includes computer instructions. When executing the computer instructions, the processor can implement the descriptions of the recommendation method for a media object in the embodiment corresponding to FIG. 3 or FIG. 15. Therefore, details are not described herein again. In addition, beneficial effects of the same method are not described herein again. For technical details that are not disclosed in the embodiment of the computer-readable storage medium in the present disclosure, refer to the descriptions of the method embodiments of the present disclosure. In

an example, the computer instructions may be deployed on one computer device for execution, or may be executed on a plurality of computer devices at one location, or may be executed on a plurality of computer devices that are distributed at a plurality of locations and that are interconnected through a communication network. The plurality of computer devices that are distributed at a plurality of locations and that are interconnected through a communication network may form a blockchain system.

[00308] In addition, the embodiments of the present disclosure further provide a computer program product or a computer program. The computer program product or the computer program may include computer instructions, and the computer instructions may be stored in a computer-readable storage medium. A processor of a computer device reads the computer instructions from the computer-readable storage medium, and the processor may execute the computer instructions, to enable the computer device to implement the descriptions of the recommendation method for a media object in the embodiment corresponding to FIG. 3 or FIG. 15. Therefore, details are not described herein again. In addition, beneficial effects of the same method are not described herein again. For technical details that are not disclosed in the embodiment of the computer program product or the computer program in the present disclosure, refer to the descriptions of the method embodiments of the present disclosure.

[00309] In the embodiments of the present disclosure, the term "module" or "unit" is a computer program having a predetermined function or a part of a computer program, operates together with other related parts to achieve a predetermined objective, and may be all or partially implemented by software, hardware (for example, a processing circuit or a memory), or a combination thereof. Similarly, one processor (or a plurality of processors or memories) may be configured to implement one or more modules or units. In addition, each module or unit may be a part of an overall module or unit including a function of the module or unit.

[00310] What is disclosed above is merely exemplary embodiments of the present disclosure, and certainly is not intended to limit the scope of the claims of the present disclosure. Therefore, equivalent variations made in accordance with the claims of the present disclosure shall fall within the scope of the present disclosure.

CLAIMS

What is claimed is:

1. A recommendation method for a media object, comprising:

obtaining target audio data of a target media object to be recommended in a target sampling window, and performing text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the target media object;

obtaining N historical text representation vectors for representing N historical media objects, and performing vector similarity calculation on the target text representation vector and each of the N historical text representation vectors respectively, to obtain N vector similarities, an interaction data amount of each of the N historical media objects being greater than an interaction data amount of the target media object, N being an integer greater than 1, each of the N historical text representation vectors being recorded after text vector representation processing is performed on a historical speech recognition text corresponding to each piece of historical audio data, each piece of historical audio data being audio data of each historical media object in a recorded historical sampling window, and a window duration of the historical sampling window being consistent with a window duration of the target sampling window;

selecting, based on the N vector similarities, K historical media objects from the N historical media objects, and establishing a similarity association relationship between the K historical media objects and the target media object, K being a positive integer less than N; and

obtaining K historical recommendation features of the K historical media objects, and determining, based on the similarity association relationship and the K historical recommendation features, a target recommendation feature for representing the target media object, and recommending the target media object based on the target recommendation feature.

2. The method according to claim 1, wherein the obtaining target audio data of the target media object to be recommended in the target sampling window comprises:

obtaining a target dynamic sampling time point and window duration of the target media object, and determining, based on the target dynamic sampling time point and the window duration, the target sampling window for sampling multimedia data of the target media object, the target dynamic sampling time point being a time point changing with a change in media data association duration of the target media object;

obtaining, from the multimedia data of the target media object based on the target sampling window, audio data with a data sampling end time point being the target dynamic sampling time point and with sampling duration being the window duration; and

using the obtained audio data as the target audio data of the target media object in the target sampling window.

3. The method according to claim 1 or 2, wherein the target dynamic sampling time point comprises a first dynamic sampling time point, the target sampling window comprises a first sampling window corresponding to the first dynamic sampling time point, and the target audio data comprises first audio data captured based on the first dynamic sampling time point and the window duration; and

the method further comprises:

obtaining a sampling interval duration for dynamic sampling, and determining, based on the first dynamic sampling time point and the sampling interval duration, a second dynamic sampling time point for sampling the multimedia data of the target media object, the second dynamic sampling time point being a next sampling time point of the first dynamic sampling time point, wherein the sampling interval duration is an interval between the first dynamic sampling time point and the second dynamic sampling time point;

determining, based on the second dynamic sampling time point and the window duration, a second sampling window for sampling the multimedia data of the target media object, the second sampling window being a next sampling window of the first sampling window, wherein the sampling interval duration is less than or equal to the window duration;

obtaining, from the multimedia data of the target media object based on the second sampling window, audio data with a data sampling end time point being the second dynamic sampling time point and with sampling duration being the window duration; and

using the obtained audio data as second audio data of the target media object in the target sampling window, updating the target dynamic sampling time point by using the second dynamic sampling time point, and updating the target audio data by using the second audio data.

4. The method according to any one of claims 1 to 3, wherein the performing text vector representation processing on the target speech recognition text corresponding to the target audio data, to obtain the target text representation vector for representing the target media object comprises:

obtaining a speech conversion understanding component and a similarity search component for

performing speech conversion processing on the target audio data, the speech conversion understanding component comprising a speech recognizer and a semantic understanding model for performing semantic extraction, and the similarity search component comprising a text representer;

performing speech recognition processing on the target audio data through the speech recognizer to obtain a speech recognition processing result;

in response to the speech recognition processing result indicating that the target speech recognition text of the target audio data is recognized, performing text semantic understanding processing on the target speech recognition text through the semantic understanding model to obtain target semantic understanding information of the target speech recognition text; and

performing vector representation on the target semantic understanding information through the text representer to obtain the target text representation vector for representing the target media object.

5. The method according to any one of claims 1 to 4, further comprising:

in response to the speech recognition processing result indicating that no target speech recognition text of the target audio data is recognized, marking the target audio data as invalid audio data.

6. The method according to any one of claims 1 to 5, further comprising:

in response to the speech recognition processing result indicating that no target speech recognition text of the target audio data is recognized, obtaining video data of the target media object from the multimedia data of the target media object based on the target sampling window, and using the obtained video data as target video data of the target media object in the target sampling window; and

performing video content recognition on a video frame in the target video data to obtain video text content for representing the target video data, and updating the target speech recognition text by using the video text content.

7. The method according to any one of claims 1 to 6, wherein the performing text semantic understanding processing on the target speech recognition text through the semantic understanding model to obtain the target semantic understanding information of the target speech recognition text comprises:

obtaining guiding prompt template data for guiding the semantic understanding model to generate model output;

inputting the guiding prompt template data and the target speech recognition text to the semantic understanding model, and performing semantic information extraction on the target speech recognition text through the semantic understanding model based on data content in the guiding prompt template data to obtain target semantic extraction information of the target speech recognition text; and

determining the target semantic understanding information of the target speech recognition text based on the target semantic extraction information.

8. The method according to any one of claims 1 to 7, wherein the obtaining guiding prompt template data for guiding the semantic understanding model to generate the model output comprises:

in response to an object type of the target media object being a video livestreaming type, using prompt template data associated with the video livestreaming type as the guiding prompt template data; or

in response to an object type of the target media object being a video-on-demand type, using prompt template data associated with the video-on-demand type as the guiding prompt template data.

9. The method according to any one of claims 1 to 8, wherein the target text representation vector is obtained by performing speech conversion processing on the target audio data through the speech conversion understanding component and the similarity search component; and

the obtaining N historical text representation vectors for representing the N historical media objects comprises:

obtaining interaction data curve distribution corresponding to a historical media object database, the interaction data curve distribution being determined based on each historical media object in the historical media object database and an interaction data amount of the historical media object;

obtaining, from the historical media object database based on the interaction data curve distribution, the N historical media objects with interaction data amounts reaching an interaction data threshold; and

obtaining the N historical text representation vectors of the N historical media objects from a historical text vector representation database associated with the N historical media objects, each historical text representation vector being obtained by performing speech conversion processing on historical audio data of each historical media object through the speech conversion understanding

component and the similarity search component.

10. The method according to any one of claims 1 to 9, wherein the selecting, based on the N vector similarities, the K historical media objects from the N historical media objects, and establishing the similarity association relationship between the K historical media objects and the target media object comprises:

sorting the N historical media objects based on the N vector similarities to obtain a historical media object list associated with the N historical media objects; and

obtaining, from the historical media object list, the K historical media objects with vector similarities meeting a similarity retrieval condition, and establishing the similarity association relationship between each of the K historical media objects and the target media object,

wherein the K historical media objects constitute a similar media object list, and each historical media object in the similar media object list has the similarity association relationship with the target media object.

11. The method according to any one of claims 1 to 10, wherein the determining, based on the similarity association relationship and K historical recommendation features when the K historical recommendation features of the K historical media objects are obtained, the target recommendation feature for representing the target media object, comprises:

when the K historical recommendation features of the K historical media objects are obtained, performing feature pooling processing on the K historical recommendation features based on the similarity association relationship to obtain a pooling processing feature associated with the K historical recommendation features; and

determining the pooling processing feature as the target recommendation feature for representing the target media object.

12. The method according to any one of claims 1 to 11, further comprising:

obtaining the K historical recommendation features of the K historical media objects from a recommendation feature database associated with the K historical media objects.

13. The method according to any one of claims 1 to 12, further comprising:

obtaining N pieces of historical interaction behavior information corresponding to the N historical media objects, one historical media object corresponding to one piece of historical interaction behavior information, and one piece of historical interaction behavior information being configured for representing item interaction details of each interaction item in one historical media

object;

using item interaction details represented by the N pieces of historical interaction behavior information as a training sample for training an initial recommendation model, and obtaining a recommendation indicator for training the initial recommendation model;

performing model training on the initial recommendation model by using the training sample and the recommendation indicator to obtain a target recommendation model, the target recommendation model being configured for extracting the N historical recommendation features of the N historical media objects, and one historical recommendation feature being obtained by performing, by the target recommendation model, interaction vector extraction on item interaction details represented by one piece of historical interaction behavior information; and

adding the N historical recommendation features extracted by the target recommendation model to a recommendation feature database.

14. The method according to any one of claims 1 to 13, wherein the target media object is a multimedia object, on a service processing platform, that has an interaction data amount less than an interaction data amount threshold, a recommendation model corresponding to the service processing platform is the target recommendation model, and the target recommendation model is obtained by performing, based on the N pieces of historical interaction behavior information corresponding to the N historical media objects, model training on an initial recommendation model deployed on the service processing platform; and

the recommending the target media object based on the target recommendation feature comprises:

when the target recommendation model is deployed and applied on the service processing platform, obtaining object association data information of a service object accessing the service processing platform;

inputting the object association data information and the target recommendation feature to the target recommendation model, and when an object vector representation feature for representing the service object is extracted from the object association data information through the target recommendation model, selecting, from the object vector representation feature, a target object vector representation feature matching the target recommendation feature; and

determining a service object corresponding to the target object vector representation feature as a target service object that is selected from the service object and that matches the target media object, and recommending the target media object to the target service object.

15. The method according to any one of claims 1 to 14, wherein the inputting the object association data information and the target recommendation feature to the target recommendation model, and when an object vector representation feature for representing the service object is extracted from the object association data information through the target recommendation model, selecting, from the object vector representation feature, the target object vector representation feature matching the target recommendation feature comprises:

inputting the object association data information and the target recommendation feature to the target recommendation model, and performing information vector representation extraction on the object association data information through the target recommendation model to obtain the object vector representation feature for representing the service object;

performing feature matching between the object vector representation feature and the target recommendation feature to obtain a feature matching result; and

in response to the feature matching result indicates that the object vector representation feature comprising an object vector representation feature matching the target recommendation feature, determining the object vector representation feature matching the target recommendation feature as the target object vector representation feature.

16. The method according to any one of claims 1 to 15, wherein a target recommendation service associated with the service processing platform comprises a livestreaming recommendation service; and

the recommending the target media object to the target service object comprises:

in response to the target media object being a livestreaming media object, using the livestreaming media object as a target recommendation object to be recommended to the target service object in the livestreaming recommendation service, and transmitting the target recommendation object to a terminal device corresponding to the target service object, to enable the terminal device to display the target recommendation object in a livestreaming recommendation list corresponding to the livestreaming recommendation service; or

in response to the target media object being a video-on-demand media object, packaging a video-on-demand link corresponding to the video-on-demand media object into a recommendation notification message, and transmitting the recommendation notification message to a terminal device corresponding to the target service object, to enable the terminal device to display, in a livestreaming recommendation list corresponding to the livestreaming recommendation service, the video-on-demand link in the recommendation notification message, the video-on-demand link being

configured for indicating the terminal device to display the video-on-demand media object in response to a trigger operation on the recommendation notification message.

17. A recommendation apparatus for a media object, comprising:

an audio data speech recognition module, configured to obtain target audio data of a target media object to be recommended in a target sampling window, and perform text vector representation processing on target speech recognition text corresponding to the target audio data, to obtain a target text representation vector for representing the target media object;

a vector similarity calculation module, configured to obtain N historical text representation vectors for representing N historical media objects, and perform vector similarity calculation on the target text representation vector and each of the N historical text representation vectors respectively, to obtain N vector similarities, an interaction data amount of each of the N historical media objects being greater than an interaction data amount of the target media object, N being an integer greater than 1, each of the N historical text representation vectors being recorded after text vector representation processing is performed on a historical speech recognition text corresponding to each piece of historical audio data, each piece of historical audio data being audio data of each historical media object in a recorded historical sampling window, and a window duration of the historical sampling window being consistent with a window duration of the target sampling window;

a similarity association relationship establishment module, configured to select, based on the N vector similarities, K historical media objects from the N historical media objects, and establish a similarity association relationship between K historical media objects and the target media object, K being a positive integer less than N; and

a target recommendation feature determining module, configured to obtain K historical recommendation features of the K historical media objects, and determine, based on the similarity association relationship and the K historical recommendation features, a target recommendation feature for representing the target media object, and recommend the target media object based on the target recommendation feature.

18. A computer device, comprising a memory and a processor,

the memory being connected to the processor, the memory being configured to store a computer program, and the processor being configured to invoke the computer program, to enable the computer device to perform the method according to any one of claims 1 to 16.

19. A computer-readable storage medium, the computer-readable storage medium having a computer program stored therein, and the computer program being adapted to be loaded and

executed by a processor, to enable a computer device with the processor to perform the method according to any one of claims 1 to 16.

20. A computer program product, comprising a computer program/instructions, the computer program/instructions, when executed by a processor, implementing the method according to any one of claims 1 to 16.