

MODEL TRAINING METHOD, APPARATUS, AND DEVICE, AND DATA SIMILARITY DETERMINING METHOD, APPARATUS, AND DEVICE

Technical Field

5 The present application relates to the field of computer technologies, and more particularly to a model training method, apparatus, and device, and a data similarity determining method, apparatus, and device.

Technical Background

As a novel identity verification method, face recognition has created new risks while
10 providing convenience to users. For a plurality of users having very similar looks (such as twins), it is difficult to effectively distinguish different users through face recognition, which is very likely to cause risks of account mis-registration and misappropriation of account funds due to the inability to correctly identify the users. As the most typical known case involving very similar looks, twins, especially identical twins, are closely related to each
15 other and are very likely to engage in behaviors associated with the above risks. How to determine user data of twins from a large amount of data has become an important issue to be solved.

Generally, based on a supervised machine learning method, a recognition model is constructed using pre-selected sample data. Specifically, an investigator conducts a social
20 survey through questionnaires, prize-winning questions, or manual observations, collects user data, and obtains and labels an association or twin relationship between users through manual observation or by asking people to be investigated. Based on the manually labeled association or twin relationship, the identification model is constructed by using the corresponding user data as sample data.

However, the above-mentioned identification model constructed by using the supervised machine learning method requires manual labeling of the sample data, and the manual labeling process consumes a large amount of manpower resources, and also consumes a large amount of time for labeling, thus making the model training inefficient and leading to high resource consumption.

Summary of the Invention

An object of the embodiments of the present application is to provide a model training method, apparatus, and device, and a data similarity determining method, apparatus, and device, to implement rapid training of a model, improve the model training efficiency, and reduce the resource consumption.

To solve the above-mentioned technical problem, the embodiments of the present application are implemented as follows:

The embodiments of the present application provide a model training method, including:

acquiring a plurality of user data pairs, where data fields of two sets of user data in each user data pair have an identical part;

acquiring a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair;

determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model; and training the classification model based on the sample data to obtain a similarity classification model.

Optionally, the acquiring a user similarity corresponding to each user data pair includes:

acquiring biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and

determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair.

Optionally, the biological feature includes a facial image feature;

the acquiring biological features of users corresponding to a first user data pair includes:

5 acquiring facial images of the users corresponding to the first user data pair; and

performing feature extraction on the facial images to obtain facial image features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair includes:

10 determining the user similarity corresponding to the first user data pair according to the facial image features of the users corresponding to the first user data pair.

Optionally, the biological feature includes a speech feature;

the acquiring biological features of users corresponding to a first user data pair includes:

acquiring speech data of the users corresponding to the first user data pair; and

15 performing feature extraction on the speech data to obtain speech features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair includes:

20 determining the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

Optionally, the determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a classification model includes:

performing feature extraction on each user data pair in the plurality of user data pairs to

25 obtain associated user features between the two sets of user data in each user data pair; and

determining, according to the associated user features between the user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

5 Optionally, the determining, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model includes:

selecting positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold; and

10 using the positive sample features and the negative sample features as the sample data for training the classification model.

Optionally, the user feature includes a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, wherein the household registration dimension feature includes a feature of user identity information, the name dimension feature includes a feature of user name information and a feature of a degree of scarcity of a user surname, and the social feature includes a feature of social relationship information of a user.

Optionally, the positive sample features include the same quantity of features as the negative sample features.

20 Optionally, the similarity classification model is a binary classifier model.

The embodiments of the present application further provide a data similarity determining method, including:

acquiring a to-be-detected user data pair;

performing feature extraction on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features; and

determining a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and a pre-trained similarity classification model.

Optionally, the method further includes:

- 5 determining to-be-detected users corresponding to the to-be-detected user data pair as twins if the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair is greater than a predetermined similarity classification threshold.

The embodiments of the present application provide a model training apparatus,
10 including:

a data acquiring module configured to acquire a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;

a similarity acquiring module configured to acquire a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding
15 to the two sets of user data in each user data pair;

a sample data determining module configured to determine, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model; and

a model training module configured to train the classification model based on the sample
20 data to obtain a similarity classification model.

Optionally, the similarity acquiring module includes:

a biological feature acquiring unit configured to acquire biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and

a similarity acquiring unit configured to determine a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair.

Optionally, the biological feature includes a facial image feature;

5 the biological feature acquiring unit is configured to acquire facial images of the users corresponding to the first user data pair; and perform feature extraction on the facial images to obtain facial image features; and

correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the facial image features of
10 the users corresponding to the first user data pair.

Optionally, the biological feature includes a speech feature;

the biological feature acquiring unit is configured to acquire speech data of the users corresponding to the first user data pair; and perform feature extraction on the speech data to obtain speech features; and

15 correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

Optionally, the sample data determining module includes:

a feature extraction unit configured to perform feature extraction on each user data pair
20 in the plurality of user data pairs to obtain associated user features between the two sets of user data in each user data pair; and

a sample data determining unit configured to determine, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

Optionally, the sample data determining unit is configured to select positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold; and use the positive sample features and the negative sample features as the sample data for training the classification model.

Optionally, the user feature includes a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, wherein the household registration dimension feature includes a feature of user identity information, the name dimension feature includes a feature of user name information and a feature of a degree of scarcity of a user surname, and the social feature includes a feature of social relationship information of a user.

Optionally, the positive sample features include the same quantity of features as the negative sample features.

Optionally, the similarity classification model is a binary classifier model.

The embodiments of the present application further provide a data similarity determining apparatus, the apparatus including:

a to-be-detected data acquiring module configured to acquire a to-be-detected user data pair;

a feature extraction module configured to perform feature extraction on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features; and

a similarity determining module configured to determine a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and a pre-trained similarity classification model.

Optionally, the apparatus further includes:

a similarity classification module configured to determine to-be-detected users corresponding to the to-be-detected user data pair as twins if the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair is greater than a predetermined similarity classification threshold.

5 The embodiments of the present application provide a model training device, including:
a processor; and

a memory configured to store computer executable instructions, wherein when being executed, the computer executable instructions cause the processor to execute the following operations:

10 acquiring a plurality of user data pairs, where data fields of two sets of user data in each user data pair have an identical part;

acquiring a user similarity corresponding to each user data pair, where the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair;

determining, according to the user similarity corresponding to each user data pair and
15 the plurality of user data pairs, sample data for training a preset classification model; and
training the classification model based on the sample data to obtain a similarity classification model.

The embodiments of the present application provide a data similarity determining device, including:

20 a processor; and

a memory configured to store computer executable instructions, where when being executed, the computer executable instructions cause the processor to execute the following operations:

acquiring a to-be-detected user data pair;

performing feature extraction on each set of to-be-detected user data in the to-be-detected user data pair, to obtain to-be-detected user features; and

determining a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features
5 and a pre-trained similarity classification model.

As can be seen from the technical solutions provided by the embodiments of the present application, in the embodiments of the present application, a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for
10 training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity classification model. In this way, a plurality of user data pairs is obtained through the same data field, and an
15 association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

20 **Brief Description of the Drawings**

To describe the technical solutions in the embodiments of the present application or the prior art more clearly, the drawings that need to be used in the description of the embodiments or the prior art are briefly introduced. It would be obvious that the drawings in the following description are merely some embodiments of the present application, and

those of ordinary skill in the art may further obtain other drawings according to these drawings without creative efforts.

FIG. 1 shows an embodiment of a model training method according to the present application;

5 FIG. 2 shows an embodiment of a data similarity determining method according to the present application;

FIG. 3 is a schematic diagram of an interface of a detection application according to the present application;

10 FIG. 4 shows an embodiment of a data similarity determining method according to the present application;

FIG. 5 is a schematic diagram of processing logic of a data similarity determining process according to the present application;

FIG. 6 shows an embodiment of a model training apparatus according to the present application;

15 FIG. 7 shows an embodiment of a data similarity determining apparatus according to the present application;

FIG. 8 shows an embodiment of a model training device according to the present application; and

20 FIG. 9 shows an embodiment of a data similarity determining device according to the present application.

Detailed Description

Embodiments of the present application provide a model training method, apparatus, and device, and a data similarity determining method, apparatus, and device.

25 To enable those skilled in the art to better understand the technical solutions of the present application, the technical solutions in the embodiments of the present application will be

clearly and fully described below with reference to the accompanying drawings in the embodiments of the present application. It would be obvious that the described embodiments are merely some rather than all embodiments of the present application. All other embodiments obtained by those of ordinary skill in the art based on the embodiments of the present application shall fall within the protection scope of the present application.

Embodiment 1

As shown in FIG. 1, the embodiment of the present application provides a model training method. The method may be executed by a terminal device or a server. The terminal device may be a personal computer or the like. The server may be an independent single server or may be a server cluster formed by a plurality of servers. The embodiment of the present application is described in detail by using an example in which the method is executed by a server to improve the model training efficiency. The method specifically may include the following steps:

In step S102, a plurality of user data pairs are acquired, wherein data fields of two sets of user data in each user data pair have an identical part.

Each user data pair may include user data of a plurality of different users. For example, the plurality of user data pairs includes a user data pair A and a user data pair B. The user data pair A includes user data 1 and user data 2, and the user data pair B includes user data 3 and user data 4, etc. The user data may be data related to a user, for example, identity information such as a name, age, height, address, identity card number, and social security card number of the user, and may also include information such as an interest, a purchased commodity, travel, etc. of the user. The data field may be a field or character capable of representing identities of users corresponding to the two sets of different user data in the user data pair as well as an association between the users, for example, a surname, a preset quantity of digits in the identity card number (for example, the first 14 digits in the identity

card number), a social security card number or other identity numbers capable of determining the user identity or information, and so on.

In an implementation, the user data may be acquired in various manners. For example, the user data may be purchased from different users; or may be information entered by the user when registering with a website or application, for example, information entered when the user registers with Alipay®; or may be user data actively uploaded by the user. The specific manner in which the user data is acquired is not limited in the embodiments of the present application. After the user data is acquired, data fields included in the acquired user data may be compared to find user data having data fields that share an identical part. The user data having data fields that share an identical part may be grouped together to form a user data pair. By means of the above method, a plurality of user data pairs may be obtained, and data fields of user data in each user data pair have an identical part.

For example, in practical applications, to reduce the amount of computation and improve the processing efficiency as much as possible, the data fields may be set as identity card numbers and surnames, and information such as the identity card number and the name of the user may be searched in the user data considering that one or more digits, for example, the first 14 digits of the identity card number, in the identity card number can represent the relationship between two users. In the embodiments of the present application, as an example, the first 14 digits of the identity card number are used as a basis for determining whether data fields have an identical part. Specifically, the first 14 digits of the identity card number and surname of each user may be acquired, and the first 14 digits of the identity card numbers and the surnames of different users may be compared. Two sets of user data that have the same surname and the same first 14 digits of the identity card number may be grouped into one user data pair. Specifically, the user data pair may be stored in the form of a user pair,

for example, {identity card number of user 1, identity card number of user 2, name of user 1, name of user 2, other data of user 1, other data of user 2}, or the like.

It should be noted that, the data fields of the two sets of user data having an identical part may be construed as that some contents in the data fields are identical, for example, the first 14 digits in 18-digit identity card numbers, or may be construed as that all contents of the data fields are identical.

In step S104, a user similarity corresponding to each user data pair is acquired, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair.

The user similarity may be used for representing the degree of similarity between a plurality of users, for example, 99% or 50%. In practical applications, the user similarity may also be represented in other manners. For example, the user similarity may also be represented by twins and non-twins, or by identical twins and fraternal twins.

In an implementation, the main purpose of this embodiment is to train a classification model, which requires sample data for training the classification model as well as a user similarity corresponding to the sample data. The user similarity may be prestored in a server or a terminal device. The user similarity may be determined in various manners. One optional processing method is provided below, and the following contents may be referred to for details: images of users may be acquired in advance, wherein the images may be uploaded by the users when registering with an application or website, and the users may be users corresponding to the two sets of user data included in each user data pair. The images in each user data pair may be compared, and through the comparison of the images, the similarity between the users corresponding to the two sets of user data included in the user data pair may be calculated. During the image comparison process, processing methods such as image

preprocessing, image feature extraction, and image feature comparison may be used, which are not limited in the embodiments of the present application.

In step S106, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model is
5 determined.

The classification model may be any classification model, such as a naive Bayesian classification model, a logistic regression classification model, a decision tree classification model, or a support vector machine classification model. In the embodiments of the present application, considering that the classification model is used only for determining whether
10 two different users are similar, the classification model may be a binary classification model. The sample data may be data used for training the classification model. The sample data may be the two sets of user data in the user data pair, and may also be data obtained after the above user data is processed in a certain manner. For example, feature extraction is performed on the above user data to obtain a corresponding user feature, and data of the user
15 feature may be used as the sample data.

In an implementation, a similarity threshold of, for example, 80% or 70%, may be set in advance. The user similarity corresponding to each user data pair may be respectively compared with the similarity threshold. The user data pairs corresponding to the user similarities greater than the similarity threshold may be grouped into one set, the user data
20 pairs corresponding to the user similarities less than the similarity threshold may be grouped into one set, a predetermined quantity (for example, 40000 or 50000) of user data pairs may be selected from each of the above two sets, and the selected user data pairs are used as the sample data for training the preset classification model.

It should be noted that, the sample data for training the preset classification model may
25 be selected in various other manners than the above manner. For example, features of the

two sets of user data included in each user data pair are extracted to obtain corresponding user features, and then, the user features may be grouped into the above two sets according to the user similarity corresponding to each user data pair and the similarity threshold. Data of the two sets of user features may be used as the sample data for training the preset
5 classification model.

In step S108, the classification model is trained based on the sample data to obtain a similarity classification model.

The similarity classification model may be a model used for determining the degree of similarity between different users.

10 In an implementation, based on the case in which the above selected user data pairs are used as the sample data for training the preset classification model, feature extraction may be performed on the two sets of user data in each of the selected user data pairs to obtain a corresponding user feature, and then the user features of each user data pair in the sample data may be input to the classification model for calculation. After the calculation, a
15 calculation result may be output. The calculation result may be compared with the user similarity corresponding to the corresponding user data pair to determine whether the two are the same. If the two are not the same, a related parameter of the classification model may be changed, then the user features of the user data pair are input to the modified classification model for calculation, and it is determined whether the calculation result is the same as the
20 user similarity. The procedure is repeated until the two are the same. If the two are the same, the above processing procedure may be performed on the next selected user data pair. Finally, if the calculation result obtained after the user features of each user data pair are input to the classification model is the same as the user similarity corresponding to the corresponding user data pair, the obtained classification model is the similarity classification model.

By means of the above method, the similarity classification model may be obtained. For use of the similarity classification model, reference can be made to the following related contents:

As shown in FIG. 2, the embodiment of the present application provides a similarity
5 determining method. The method may be executed by a terminal device or a server. The terminal device may be a personal computer or the like. The server may be an independent single server or may be a server cluster formed by a plurality of servers. The method specifically may include the following steps:

In step S202, a to-be-detected user data pair is acquired.

10 The to-be-detected user data pair may be a user data pair formed by user data of two users to be detected.

In an implementation, to detect the similarity between two different users, a corresponding detection application may be set. As shown in FIG. 3, the detection application may include a button for uploading data. When the similarity between two
15 different users needs to be detected, the button for uploading data may be tapped. The detection application may pop up a prompt box for uploading data. A data uploader may input data of the to-be-detected user data pair in the prompt box, and then tap a confirmation button in the prompt box when the input is completed. The detection application may acquire the to-be-detected user data pair input by the data uploader. The detection application may
20 be installed on a terminal device or may be installed on a server. If the similarity determining method provided by the embodiment of the present application is executed by a server and the detection application is installed on a terminal device, the detection application may send the to-be-detected user data pair to the server after acquiring the to-be-detected user data pair, so that the server may acquire the to-be-detected user data pair. If the detection application

is installed on the server, the server may directly acquire the to-be-detected user data pair from the detection application.

In step S204, feature extraction is performed on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features.

- 5 The to-be-detected user features may be a feature of the user data of the user to be detected.

In an implementation, each set of to-be-detected user data in the to-be-detected user data pair may be acquired. For any set of to-be-detected user data, a corresponding feature may be extracted from the to-be-detected user data by using a preset feature extraction algorithm,
10 and the extracted feature may be used as a to-be-detected user feature corresponding to the to-be-detected user data. By means of the above method, the to-be-detected user feature corresponding to each set of to-be-detected user data in the to-be-detected user data pair may be obtained.

It should be noted that the feature extraction algorithm may be any algorithm capable of
15 extracting a predetermined feature from the user data, and specifically may be set according to actual situations.

In step S206, a similarity is determined between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and a pre-trained similarity classification model.

20 In an implementation, the to-be-detected user feature obtained through step S204 may be input to the similarity classification model obtained through step S102 to step S108 for calculation. The result output from the similarity classification model may be the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair.

It should be noted that, in practical applications, the direct output result of the similarity classification model may be presented in percentage, for example, 90% or 40%. To make the output result more intuitive for the user, the direct output result of the similarity classification model may further be set according to actual situations, such as when identical twins and non-identical twins need to be distinguished, or when identical twins and fraternal twins need to be distinguished. In view of the above case, a classification threshold may be set. If the direct output result is greater than the classification threshold, it is determined that the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair are identical twins; otherwise, the users are determined as non-identical twins or fraternal twins. In this way, the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair may be rapidly determined according to the pre-trained similarity classification model, thereby improving the efficiency of determining the similarity between the users.

It should be noted that both the user data pair and the to-be-detected user data pair in the foregoing descriptions include two sets of user data, and in practical applications, the model training method and the similarity determining method provided by the present application may also be applied to a user data combination and a to-be-detected user data combination including more than two sets of user data. For specific implementations, reference can be made to the related contents in the embodiments of the present application, and the details are not described herein again.

The embodiment of the present application provides a model training method and similarity determining method, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the

sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity classification model. In this way, a plurality of user data pairs is obtained only through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

Embodiment 2

As shown in FIG. 4, the embodiment of the present application provides a data similarity determining method. The method may be executed by a server, or jointly by a terminal device and a server. The terminal device may be a personal computer or the like. The server may be an independent single server or may be a server cluster formed by a plurality of servers. In the embodiment of the present application, to improve the model training efficiency, a detailed description is made by using an example where the method is executed by a server. For the case that the method is implemented jointly by a terminal device and a server, reference can be made to the following related contents, and the details are not described herein again. The method specifically includes the following contents:

At present, as a novel user identity verification method, face recognition has created new risks while providing convenience to users. For current face recognition technologies, an image of a user is captured on site and compared with a user image of the user that is prestored in a database of a face recognition system, and if the value obtained through comparison reaches a predetermined threshold, it is determined that the user is the user corresponding to the prestored user image, thus verifying the identity of the user. However,

it is difficult to effectively verify the identities of users having very similar faces by using the above method, which is very likely to cause account mis-registration and misappropriation of account funds due to the inability to verify the identity.

As the most typical known case involving very similar looks, twins, especially identical twins, are closely related to each other and are very likely to cause negative public opinions. If there is a list including as many twins users as possible, a special face recognition coping strategy may be designed for these users to prevent the above risk. Therefore, a model for effectively identifying twins may be constructed to output a twin list for monitoring face recognition behavior of these users while ensuring high accuracy, thus achieving risk control.

For the implementation of constructing the model for effectively identifying twins, reference can be made to the model training method provided by step S402 to step S412 below, and specific contents are as follows:

In step S402, a plurality of user data pairs are acquired, wherein data fields of two sets of user data in each user data pair have an identical part.

In an implementation, considering that twins generally have the same surname and the same first 14 digits of the identity card numbers, surnames and the first 14 digits of identity card numbers may be used as data fields for selecting the user data pair. For a specific implementation of step S402, reference can be made to the related content of step S102 in Embodiment 1, and the details are not described herein again.

It should be noted that, the processing of selecting the user data pair is implemented based on the surnames and the first 14 digits of the identity card numbers. In another embodiment of the present application, the processing of selecting the user data pair may also be implemented based on other information, for example, the surnames and the social security card numbers, or the first 14 digits of the identity card numbers and the social security card numbers, and is not limited in the embodiments of the present application.

Considering that the degree of similarity between the users corresponding to the two sets of user data in the user data pair needs to be determined during model training, the following provides a related processing manner, and reference can be made specifically to step S404 and step S406 below.

5 In step S404, biological features of users corresponding to a first user data pair are acquired, wherein the first user data pair is any user data pair in the plurality of user data pairs.

The biological features may be physiological and behavioral features of a human body, such as a fingerprint feature, iris feature, facial feature, DNA, or other physiological features,
10 or a voiceprint feature, handwriting feature, keystroke habit, or other behavioral features.

In an implementation, after the plurality of user data pairs is acquired through step S402 above, a user data pair (that is, the first user data pair) may be arbitrarily selected from the plurality of user data pairs. When the user uses a terminal device to log in to a server for registration, the user may upload one or more of the above biological features of the user to
15 the server. The server may store the biological feature and an identifier of the user in an associated manner. The identifier of the user may be a username or a name of the user input by the user during registration. The above information stored in an associated manner in the server may be as shown in Table 1.

Table 1

identifier of the user	biological feature
user 1	biological feature A
user 2	biological feature B
user 3	biological feature C

20 After selecting the first user data pair, the server may extract identifiers of users included in the first user data pair, and then may acquire corresponding biological features according

to the identifiers of the users, thus obtaining the biological features of the users corresponding to the first user data pair. For example, the identifiers of the users included in the first user data pair are user 2 and user 3, and by querying the corresponding relationships in the above table, it may be determined that the user 2 corresponds to a biological feature B, and the user 3 corresponds to a biological feature C, that is, the biological features of the users corresponding to the first user data pair are the biological feature B and the biological feature C.

In step S406, a user similarity corresponding to the first user data pair is determined according to the biological features of the users corresponding to the first user data pair.

10 In an implementation, after the biological features of the users corresponding to the first user data pair are obtained through step S404 above, similarity calculation may be respectively performed for the obtained biological features, so as to determine the degree of similarity between two corresponding users (that is, the user similarity). The similarity calculation may be implemented in various manners, for example, according to a Euclidean distance between feature vectors, and is not limited in the embodiments of the present application.

It should be noted that a threshold may be set for determining whether users are similar. For example, the threshold is set to 70. When the user similarity corresponding to two biological features is greater than 70, it is determined that the users corresponding to the two sets of user data in the first user data pair are similar; when the user similarity corresponding to two biological features is less than 70, it is determined that the users corresponding to the two sets of user data in the first user data pair are not similar.

By means of the above method, the above processing procedure may be performed on other user data pairs in addition to the first user data pair in the plurality of user data pairs,

so as to obtain the user similarity corresponding to each user data pair in the plurality of user data pairs.

In step S404 and step S406 above, the user similarity is determined according to the biological features of the users. In practical applications, the user similarity may specifically be determined in various manners. Step S404 and step S406 are specifically described below by using an example where the biological features are facial features, and reference can be made to the following step 1 and step 2 for details.

In step 1, facial images of the users corresponding to the first user data pair are acquired, wherein the first user data pair is any user data pair in the plurality of user data pairs.

In an implementation, after the plurality of user data pairs are acquired through step S402 above, one user data pair (that is, the first user data pair) may be arbitrarily selected from the plurality of user data pairs. When the user uses a terminal device to log in to a server for registration, the user may upload an image including the face of the user to the server. The server may store the image and an identifier of the user in an associated manner. The identifier of the user may be a username or a name of the user input by the user during registration. The above information stored in an associated manner in the server may be as shown in Table 2.

Table 2

identifier of user	image including the face of the user
user 1	image A
user 2	image B
user 3	image C

After acquiring the first user data pair, the server may extract identifiers of users included in the first user data pair, and then may acquire corresponding images according to the identifiers of the users, thus obtaining the facial images of the users corresponding to the

first user data pair. For example, the identifiers of the users included in the first user data pair are user 2 and user 3, and by querying the corresponding relationships in the above table, it may be determined that the image including the face of the user and corresponding to the user 2 is an image B, and the image including the face of the user and corresponding to the user 3 is an image C, that is, the facial images of the users corresponding to the first user data pair are the image B and the image C.

In step 2, feature extraction is performed on the facial images to obtain facial image features, and the user similarity corresponding to the first user data pair is determined according to the facial image features of the users corresponding to the first user data pair.

In an implementation, after the facial images of the users corresponding to the first user data pair are obtained through step 1, feature extraction may be respectively performed on the obtained facial images to obtain corresponding facial image features, and a corresponding feature vector is obtained based on an extracted feature of each facial image; then a Euclidean distance between feature vectors of any two facial images may be calculated, and according to the value of the Euclidean distance between the feature vectors, the degree of similarity between the two corresponding users (that is, the user similarity) may be determined. The larger the value of the Euclidean distance between feature vectors is, the lower the user similarity is; the smaller the value of the Euclidean distance between feature vectors is, the higher the user similarity is.

It should be noted that, for facial images, two facial images may either be similar or not. Therefore, a threshold may be set for determining whether images are similar. For example, the threshold is set to 70. When the user similarity corresponding to two facial images is greater than 70, it is determined that the users corresponding to the two sets of user data in the first user data pair are similar; when the user similarity corresponding to two facial

images is less than 70, it is determined that the users corresponding to the two sets of user data in the first user data pair are not similar.

For example, based on the example of step 1, feature extraction is respectively performed on the image B and the image C, and corresponding feature vectors are constructed
5 respectively according to the extracted features to obtain a feature vector of the image B and a feature vector of the image C. A Euclidean distance between the feature vector of the image B and the feature vector of the image C is calculated, and the user similarity between the user 2 and the user 3 is determined according to the value of the obtained Euclidean distance.

By means of the above method, the above processing procedure may be performed on
10 other user data pairs in addition to the first user data pair in the plurality of user data pairs, so as to obtain the user similarity corresponding to each user data pair in the plurality of user data pairs.

In addition, for processing of step S404 and step S406 above, the following further provides an optional processing manner, and reference can be made to the following step 1
15 and step 2 for details.

In step 1, speech data of the users corresponding to the first user data pair is acquired, wherein the first user data pair is any user data pair in the plurality of user data pairs.

In an implementation, after the plurality of user data pairs are acquired through step S402 above, one user data pair (that is, the first user data pair) may be arbitrarily selected from the
20 plurality of user data pairs. When the user uses a terminal device to log in to a server for registration, the user may upload speech data having a predetermined duration (for example, 3 seconds or 5 seconds) and/or including a predetermined speech content (for example, speech of one or more words or one sentence) to the server. The server may store the speech data and an identifier of the user in an associated manner. After selecting the first user data
25 pair, the server may respectively extract the identifiers of the users included in the first user

data pair, and then acquire corresponding speech data according to the identifiers of the users, thus obtaining the speech data of the users corresponding to the first user data pair.

In step 2, feature extraction is performed on the speech data to obtain speech features, and the user similarity corresponding to the first user data pair is determined according to
5 the speech features of the users corresponding to the first user data pair.

In an implementation, after the speech data of the users corresponding to the first user data pair are obtained through step 1 above, feature extraction may be respectively performed on the obtained speech data, and based on an extracted feature of each piece of speech data, the degree of similarity between the two corresponding users (that is, the user similarity)
10 may be determined. For a specific implementation, reference can be made to the related content in step S406 above. Alternatively, the user similarity may be determined through one-by-one comparison of features; or speech spectrum analysis may be performed for any two pieces of speech data to determine the user similarity. By means of the above method, the above processing procedure may be performed on other user data pairs in addition to the
15 first user data pair in the plurality of user data pairs, so as to obtain the user similarity corresponding to each user data pair in the plurality of user data pairs.

In step S408, feature extraction is performed on each user data pair in the plurality of user data pairs to obtain associated user features between the two sets of user data in each user data pair.

20 In an implementation, a user data pair (which may be referred to as a third user data pair) may be arbitrarily selected from the plurality of user data pairs, feature extraction may be respectively performed on two sets of different user data in the third user data pair. For example, the third user data pair includes user data 1 and user data 2, and feature extraction may be respectively performed on the user data 1 and the user data 2. Then, the associated
25 user features between the two sets of user data in the third user data pair may be obtained by

comparing features extracted from different user data. By means of the above method, the above processing procedure may be performed on other user data pairs in addition to the third user data pair in the plurality of user data pairs, so as to obtain the associated user features between the two sets of user data in each user data pair.

5 In practical applications, the user feature may include, but is not limited to, a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, or the like. The household registration dimension feature may include a feature of user identity information. The household registration dimension feature is mainly based on the household registration management system of China. The identity card information
10 included in the household registration includes the date of birth and the household registration place, and the household registration includes the parents' names and citizen's address. However, due to historical and other reasons, some citizens' registration information is not the same as the actual situation. For example, the registered date of birth may be earlier than the real date, two children may respectively follow the parents' surnames, or even the
15 divorce of the parents leads to the separation of the household registration. Therefore, the household registration dimension feature may serve as a reference for determining whether the two users are twins. In this way, the association between different users is determined depending on features such as whether different users corresponding to the user data pair have the same date of birth, the same household registration place, the same parents, or the
20 same current address.

 The name dimension feature includes a feature of user name information and a feature of a degree of scarcity of a user surname. For the name dimension feature, based on the Nature Language Processing (NLP) theory and social experience, generally, if the names of two people look alike, for example, Zhang Jinlong and Zhang Jinhu, or have a certain
25 semantic relationship, for example, Zhang Meimei and Zhang Lili, it is considered that there

is an association between the two. In the embodiment of the present application, the relationship between the names of the two users may be assessed by using a dictionary, and the user's registered personal information and demographic data are used to calculate the degree of scarcity of the surname as a feature. In this way, the association between different users is determined depending on features such as whether different users corresponding to the user data pair have the same surname or have the same length of name, the degree of synonym of the names, whether the combination of the names is a phrase, the degree of scarcity of the surname, or the like.

The social feature includes a feature of social relationship information of a user. The social feature may be obtained by extracting a social relationship of the user data pair based on big data. Generally, twins would interact with each other frequently and have a highly overlapping social relationship, for example, same relatives or even classmates. In the embodiment of the present application, the user data pair is associated based on a relationship network formed by personal information of users stored in the server and existing data, address books, or the like, to obtain corresponding features. In this way, the association between different users corresponding to the user data pair is determined depending on features such as whether the different users follow each other in a social networking application, whether the different users have transferred funds between each other, whether the different users have saved contact information of the other party to the address book, whether the different users have marked a specific appellation for the other party in the address book, a quantity of common contacts between their address books, or the like.

In addition, considering that twins have many hobbies and shopping preferences in common and may travel together, the user feature may further include features of e-commerce, tourism, entertainment, and other dimensions. In the embodiment of the present application, data related to the features of e-commerce, tourism, entertainment and other

dimensions may be acquired from a predetermined database or website. In this way, the association between different users corresponding to the user data pair is determined depending on features such as a quantity of common shopping records between the different users, whether they have traveled together, whether they have checked in at a hotel at the same time, a similarity between their shopping preferences, whether they have the same delivery address, or the like.

It should be noted that the processing of determining the user similarity (including step S404 and step S406) and the processing of feature extraction (that is, step S408) above are executed in a chronological order. In practical applications, the processing of determining the user similarity and the processing of feature extraction may also be executed at the same time or in a reversed order, which is not limited in the embodiments of the present application.

In step S410, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model is determined.

In an implementation, a threshold may be set in advance. According to the threshold, user data pairs corresponding to the user similarities greater than the threshold may be selected from the plurality of user data pairs. The associated user features between the two sets of user data in each of the selected user data pairs may be used as user features for training the classification model. The selected user features and the user similarities corresponding to the selected user data pairs may be determined as the sample data for training the classification model.

The processing of step S410 may be implemented in various other manners in addition to the above manner. The following further provides an optional processing manner, specifically including the following step 1 and step 2:

In step 1, positive sample features and negative sample features are selected from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold.

In an implementation, based on the common sense that identical twins have highly similar looks, the same date of birth and place of birth, and generally the same surname, the user similarity between the facial images of the two users is calculated, so as to determine whether the two users are identical twins. Specifically, a similarity threshold of, for example, 80% or 70%, may be set in advance. User data pairs corresponding to the user similarities greater than the similarity threshold may be determined as user data pairs of identical twins, and user data pairs corresponding to the user similarities less than the similarity threshold may be determined as user data pairs of non-identical twins. Meanwhile, because identical twins and fraternal twins basically have the same features except in looks, user features corresponding to the user data pairs of identical twins may be used as positive sample features of the similarity classification model, and user features corresponding to the user data pairs of non-identical twins (including fraternal twins and non-twins) may be used as negative sample features of the similarity classification model.

It should be noted that, the negative sample features do not mean that features included therein are all user features of fraternal twins. In practical applications, the user features of fraternal twins may account for an extremely small portion in the negative sample features, or the negative sample features may include a small number of positive sample features, which does not affect the training of the classification model but will improve the robustness of the similarity classification model.

In addition, the positive sample features may include the same quantity of features as the negative sample features. For example, 10000 user data pairs corresponding to the user similarities less than 10% are selected from the plurality of user data pairs, 10000 user data

pairs corresponding to the user similarities greater than 10% and less than 20% are selected from the plurality of user data pairs, 10000 user data pairs corresponding to the user similarities greater than 20% and less than 30% are selected from the plurality of user data pairs, 10000 user data pairs corresponding to the user similarities greater than 30% and less than 40% are selected from the plurality of user data pairs, and 10000 user data pairs corresponding to the user similarities greater than 40% and less than 50% are selected from the plurality of user data pairs. User features of the above 50000 user data pairs are used as the negative sample features. 40000 user data pairs corresponding to the user similarities greater than 80% and less than 90% are selected from the plurality of user data pairs, and 10000 user data pairs corresponding to the user similarities greater than 90% and less than 100% are selected from the plurality of user data pairs. User features of the above 50000 user data pairs are used as the positive sample features.

In step 2, the positive sample features and the negative sample features are used as the sample data for training the classification model.

In an implementation, data of the user features and the corresponding user similarities may be combined, and the combined data may be used as the sample data for training the classification model.

In step S412, the classification model is trained based on the sample data to obtain a similarity classification model.

Because the main purpose of the classification model is to identify twins, the similarity classification model may be a binary classifier model, specifically a Gradient Boosting Decision Tree (GBDT) binary classifier model, in order to make the embodiments of the present application simple and feasible.

In an implementation, positive sample features may be respectively input to the classification model for calculation. The obtained calculation result may be compared with

the user similarity corresponding to the positive sample feature. If the two match each other, the next positive sample feature or negative sample feature may be selected and input to the classification model for calculation. The obtained calculation result continues to be compared with the user similarity corresponding to the positive sample feature. If the two
5 do not match, a related parameter of the classification model may be adjusted, then the positive sample feature is input to the adjusted classification model for calculation, and the obtained calculation result is compared with the user similarity corresponding to the positive sample feature again. The procedure is repeated until the two match each other. By means of the above method, all the positive sample features and all the negative sample features
10 may be input to the classification model for calculation, thus training the classification model. The final classification model obtained through training may be used as the similarity classification model.

Thus, the similarity classification model is obtained through the above processing procedure. The similarity classification model may be applied to a face recognition scenario.

15 For twins users who may engage in behavior related to risks, the similarity classification model can be used for separate risk control.

After the similarity classification model is obtained, it may be determined by using the similarity classification model whether the to-be-detected users corresponding to the to-be-detected user data pair are twins, as shown in FIG. 5. For a specific implementation,
20 reference can be made to the contents of the following step S414 to step S420.

In step S414, a to-be-detected user data pair is acquired.

Contents of step S414 are the same as those of step S202 in Embodiment 1, reference can be made to the related contents of step S202 for a specific implementation of step S414, and the details are not described herein again.

In step S416, feature extraction is performed on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features.

For the process of performing feature extraction on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features in step S416, reference can be made to the related contents of step S408 above. That is, the features extracted from the to-be-detected user data include, but are not limited to, a household registration dimension feature, a name dimension feature, a social feature, an interest feature, or the like. Reference can be made to the related contents of step S408, and the details are not described herein again.

In step S418, a similarity is determined between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and the pre-trained similarity classification model.

Contents of step S418 are the same as those of step S206 in Embodiment 1, reference can be made to the related contents of step S206 for a specific implementation of step S418, and the details are not described herein again.

In step S420, to-be-detected users corresponding to the to-be-detected user data pair are determined as twins if the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair is greater than a predetermined similarity classification threshold.

In an implementation, because the output twin list affects the use of the face recognition of target users, high accuracy of the similarity classification model is desired for use. In practical applications, the similarity classification threshold may be set to a large value, for example, 95% or 97%. The to-be-detected user feature is predicted and scored by using the trained similarity classification model. The scoring process is to calculate a probability that the users corresponding to the corresponding user data pair are twins. For example, if the

probability is 80%, the score is 80; if the probability is 90%, the score is 90. The higher the score is, the higher the probability that the users corresponding to the user data pair are twins.

The embodiment of the present application provides a data similarity determining method, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined subsequently according to the similarity classification model. In this way, a plurality of user data pairs is obtained only through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

Embodiment 3

The data similarity determining method provided by the embodiments of the present application is described above. Based on the same concept, the embodiment of the present application further provides a model training apparatus, as shown in FIG. 6.

The model training apparatus may be disposed in a server. The apparatus includes: a data acquiring module 601, a similarity acquiring module 602, a sample data determining module 603, and a model training module 604, wherein:

the data acquiring module 601 is configured to acquire a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;

the similarity acquiring module 602 is configured to acquire a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair;

the sample data determining module 603 is configured to determine, according to the
5 user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model; and

the model training module 604 is configured to train the classification model based on the sample data to obtain a similarity classification model.

In the embodiment of the present application, the similarity acquiring module 602
10 includes:

a biological feature acquiring unit configured to acquire biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and

a similarity acquiring unit configured to determine a user similarity corresponding to the
15 first user data pair according to the biological features of the users corresponding to the first user data pair.

In the embodiment of the present application, the biological feature includes a facial image feature;

the biological feature acquiring unit is configured to acquire facial images of the users
20 corresponding to the first user data pair; and perform feature extraction on the facial images to obtain facial image features; and

correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the facial image features of the users corresponding to the first user data pair.

In the embodiment of the present application, the biological feature includes a speech feature;

the biological feature acquiring unit is configured to acquire speech data of the users corresponding to the first user data pair; and perform feature extraction on the speech data
5 to obtain speech features; and

correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

In the embodiment of the present application, the sample data determining module 603
10 includes:

a feature extraction unit configured to perform feature extraction on each user data pair in the plurality of user data pairs to obtain associated user features between the two sets of user data in each user data pair; and

a sample data determining unit configured to determine, according to the associated user
15 features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

In the embodiment of the present application, the sample data determining unit is configured to select positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity
20 corresponding to each user data pair and a predetermined similarity threshold; and use the positive sample features and the negative sample features as the sample data for training the classification model.

In the embodiment of the present application, the user feature includes a household registration dimension feature, a name dimension feature, a social feature, and an interest
25 feature, wherein the household registration dimension feature includes a feature of user

identity information, the name dimension feature includes a feature of user name information and a feature of a degree of scarcity of a user surname, and the social feature includes a feature of social relationship information of a user.

In the embodiment of the present application, the positive sample features include the
5 same quantity of features as the negative sample features.

In the embodiment of the present application, the similarity classification model is a binary classifier model.

The embodiment of the present application provides a model training apparatus, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each
10 user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity
15 classification model. In this way, a plurality of user data pairs is obtained only through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency
20 can be improved, and the resource consumption can be reduced.

Embodiment 4

The model training apparatus provided by the embodiments of the present application is described above. Based on the same concept, the embodiment of the present application further provides a data similarity determining apparatus, as shown in FIG. 7.

The data similarity determining apparatus includes: a to-be-detected data acquiring module 701, a feature extraction module 702, and a similarity determining module 703, wherein:

the to-be-detected data acquiring module 701 is configured to acquire a to-be-detected user data pair;

the feature extraction module 702 is configured to perform feature extraction on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features; and

the similarity determining module 703 is configured to determine a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and a pre-trained similarity classification model.

In the embodiment of the present application, the apparatus further includes:

a similarity classification module configured to determine to-be-detected users corresponding to the to-be-detected user data pair as twins if the similarity between the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair is greater than a predetermined similarity classification threshold.

The embodiment of the present application provides a data similarity determining apparatus, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity classification model. In this way, a plurality of user data pairs is obtained

only through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

Embodiment 5

Based on the same concept, the embodiment of the present application further provides a model training device, as shown in FIG. 8.

The model training device may be the server or the like provided in the foregoing embodiments.

The model training device may differ greatly depending on different configurations or performance, and may include one or more processors 801 and a memory 802. The memory 802 may store one or more storage applications or data. The memory 802 may be a transitory or permanent storage. The application stored in the memory 802 may include one or more modules (not shown), wherein each module may include a series of computer executable instructions in the model training device. Still further, the processor 801 may be configured to communicate with the memory 802, and execute, on the model training device, the series of computer executable instructions in the memory 802. The model training device may further include one or more power supplies 803, one or more wired or wireless network interfaces 804, one or more input/output interfaces 805, and one or more keyboards 806.

Specifically, in this embodiment, the model training device includes the memory and one or more programs. The one or more programs are stored in the memory. The one or more programs may include one or more modules, and each module may include a series of computer executable instructions in the model training device. The one or more processors

are configured to execute the one or more programs to execute the computer executable instructions:

acquiring a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;

5 acquiring a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair;

determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model; and

10 training the classification model based on the sample data to obtain a similarity classification model.

Optionally, when being executed, the executable instructions may further cause the processor to:

15 acquire biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and

determine a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair.

Optionally, when being executed, the executable instructions may further cause the processor to operate in the following manner:

20 the biological feature includes a facial image feature;

the acquiring biological features of users corresponding to a first user data pair includes:

acquiring facial images of the users corresponding to the first user data pair; and

performing feature extraction on the facial images to obtain facial image features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair includes:

5 determining the user similarity corresponding to the first user data pair according to the facial image features of the users corresponding to the first user data pair.

Optionally, when being executed, the executable instructions may further cause the processor to operate in the following manner:

the biological feature includes a speech feature;

the acquiring biological features of users corresponding to a first user data pair includes:

10 acquiring speech data of the users corresponding to the first user data pair; and

performing feature extraction on the speech data, to obtain speech features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair includes:

15 determining the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

Optionally, when being executed, the executable instructions may further cause the processor to:

perform feature extraction on each user data pair in the plurality of user data pairs to

20 obtain associated user features between the two sets of user data in each user data pair; and

determine, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

Optionally, when being executed, the executable instructions may further cause the
25 processor to:

select positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold; and

5 use the positive sample features and the negative sample features as the sample data for training the classification model.

Optionally, the user feature includes a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, wherein the household registration dimension feature includes a feature of user identity information, the name dimension feature includes a feature of user name information and a feature of a degree of
10 scarcity of a user surname, and the social feature includes a feature of social relationship information of a user.

Optionally, the positive sample features include the same quantity of features as the negative sample features.

Optionally, the similarity classification model is a binary classifier model.

15 The embodiment of the present application provides a model training device, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification
20 model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity classification model. In this way, a plurality of user data pairs is obtained only through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for
25 training a preset classification model, that is, the sample data can be obtained without manual

labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

Embodiment 6

Based on the same concept, the embodiment of the present application further provides
5 a data similarity determining device, as shown in FIG. 9.

The data similarity determining device may be the server, the terminal device, or the like provided in the foregoing embodiments.

The data similarity determining device may differ greatly depending on different configurations or performance, and may include one or more processors 901 and a memory
10 902. The memory 902 may store one or more storage applications or data. The memory 902 may be a transitory or permanent storage. The application stored in the memory 902 may include one or more modules (not shown), wherein each module may include a series of computer executable instructions in the data similarity determining device. Still further, the processor 901 may be configured to communicate with the memory 902, and execute, on the
15 data similarity determining device, the series of computer executable instructions in the memory 902. The data similarity determining device may further include one or more power supplies 903, one or more wired or wireless network interfaces 904, one or more input/output interfaces 905, and one or more keyboards 906.

Specifically, in this embodiment, the data similarity determining device includes the
20 memory and one or more programs. The one or more programs are stored in the memory. The one or more programs may include one or more modules, and each module may include a series of computer executable instructions in the data similarity determining device. The one or more processors are configured to execute the one or more programs to execute the following computer executable instructions:

25 acquiring a to-be-detected user data pair;

performing feature extraction on each set of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features; and

determining a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features
5 and a pre-trained similarity classification model.

Optionally, when being executed, the executable instructions may further cause the processor to:

determine to-be-detected users corresponding to the to-be-detected user data pair as twins if the similarity between the users corresponding to the two sets of to-be-detected user
10 data in the to-be-detected user data pair is greater than a predetermined similarity classification threshold.

The embodiment of the present application provides a data similarity determining device, in which a plurality of user data pairs is acquired, wherein data fields of two sets of user data in each user data pair have an identical part; a user similarity corresponding to each user data
15 pair is acquired; sample data for training a preset classification model is determined; and then the classification model is trained based on the sample data to obtain a similarity classification model, so that a similarity between users corresponding to the two sets of to-be-detected user data in a to-be-detected user data pair can be determined according to the similarity classification model. In this way, a plurality of user data pairs is obtained only
20 through the same data field, and an association between users corresponding to the two sets of user data in each user data pair is determined according to the user similarity to obtain sample data for training a preset classification model, that is, the sample data can be obtained without manual labeling, so that rapid training of a model can be implemented, the model training efficiency can be improved, and the resource consumption can be reduced.

The specific embodiments of the specification have been described above. Other embodiments are within the scope of the appended claims. In some cases, the actions or steps set forth in the claims can be performed in a different order than those described in the embodiments and can still achieve the desired results. In addition, the processes depicted in the figures are not necessarily in the shown particular order or in a continuous order to achieve the desired results. In some embodiments, multitasking and parallel processing are also possible or may be advantageous.

In the 1990s, an improvement on a technology may be obviously distinguished as an improvement on hardware (for example, an improvement on a circuit structure such as a diode, a transistor, and a switch) or an improvement on software (an improvement on a method procedure). However, with the development of technologies, improvements of many method procedures at present may be considered as direct improvements on hardware circuit structures. Almost all designers program improved method procedures into hardware circuits to obtain corresponding hardware circuit structures. Therefore, it is improper to assume that the improvement of a method procedure cannot be implemented by using a material hardware module. For example, a Programmable Logic Device (PLD) (for example, a Field Programmable Gate Array (FPGA)) is such an integrated circuit whose logic functions are determined by devices programmed by a user. Designers program by themselves to "integrate" a digital system into a piece of PLD, without the need to ask a chip manufacturer to design and manufacture a dedicated integrated circuit chip. Moreover, at present, the programming is mostly implemented by using logic compiler software, instead of manually manufacturing an integrated circuit chip. The logic compiler software is similar to a software compiler used for developing and writing a program, and original codes before compiling also need to be written using a specific programming language, which is referred to as a Hardware Description Language (HDL). There are many types of HDLs, such as Advanced

Boolean Expression Language (ABEL), Altera Hardware Description Language (AHDL), Confluence, Cornell University Programming Language (CUPL), HDCal, Java Hardware Description Language (JHDL), Lava, Lola, MyHDL, PALASM, and Ruby Hardware Description Language (RHDL), among which Very-High-Speed Integrated Circuit Hardware Description Language (VHDL) and Verilog are most commonly used now. Those skilled in the art should also know that a hardware circuit for implementing the logic method procedure may be easily obtained by slightly logically programming the method procedure using the above several hardware description languages and programming it into an integrated circuit.

A controller may be implemented in any suitable manner. For example, the controller may be in the form of, for example, a microprocessor or a processor and a computer readable medium storing a computer readable program code (for example, software or firmware) executable by the (micro)processor, a logic gate, a switch, an Application Specific Integrated Circuit (ASIC), a programmable logic controller, and an embedded micro-controller.

Examples of the controller include, but are not limited to, the following micro-controllers: ARC 625D, Atmel AT91SAM, Microchip PIC18F26K20, and Silicone Labs C8051F320. A memory controller may also be implemented as a part of control logic of a memory. Those skilled in the art also know that the controller may be implemented by using pure computer readable program codes, and in addition, the method steps may be logically programmed to enable the controller to implement the same function in a form of a logic gate, a switch, an application-specific integrated circuit, a programmable logic controller and an embedded microcontroller. Therefore, this type of controller may be considered as a hardware component, and apparatuses included therein for implementing various functions may also be considered as structures inside the hardware component. Or, the apparatuses used for

implementing various functions may even be considered as both software modules for implementing the method and structures inside the hardware component.

The system, apparatus, module or unit illustrated in the above embodiments may be specifically implemented by using a computer chip or a material object, or a product having
5 a certain function. A typical implementation device is a computer. Specifically, the computer can be, for example, a personal computer, a laptop computer, a cellular phone, a camera phone, a smart phone, a personal digital assistant, a media player, a navigation device, an email device, a game console, a tablet computer, a wearable device, or a combination of any of these devices.

10 For ease of description, when the above apparatus is described, it is divided into various units in terms of functions for respective descriptions. Definitely, when the present application is implemented, functions of the units may be implemented in the same or multiple pieces of software and/or hardware.

As will be appreciated by those skilled in the art, the embodiments of the present
15 application may be embodied as a method, a system, or a computer program product. Accordingly, the present application may use the form of an entirely hardware embodiment, an entirely software embodiment or an embodiment combining software and hardware aspects. Furthermore, the embodiments of the present application may use the form of a computer program product implemented on one or more computer-usable storage media
20 (including but not limited to magnetic disk memories, CD-ROMs, optical memories, etc.) including computer-usable program code.

The present application is described with reference to flowcharts and/or block diagrams of the method, device (system) and computer program product in the embodiments of the present application. It should be understood that computer program instructions can
25 implement each process and/or block in the flowcharts and/or block diagrams and a

combination of processes and/or blocks in the flowcharts and/or block diagrams. These computer program instructions may be provided to a general-purpose computer, a special-purpose computer, an embedded processor or a processor of another programmable data processing device to generate a machine, so that an apparatus configured to implement
5 functions specified in one or more processes in the flowcharts and/or one or more blocks in the block diagrams is generated by using instructions executed by the computer or the processor of another programmable data processing device.

These computer program instructions may also be stored in a computer readable memory that can guide a computer or another programmable data processing device to work in a
10 specified manner, so that the instructions stored in the computer readable memory generate a product including an instruction apparatus, wherein the instruction apparatus implements functions specified in one or more processes in the flowcharts and/or one or more blocks in the block diagrams.

These computer program instructions may also be loaded into a computer or another
15 programmable data processing device, so that a series of operation steps are performed on the computer or another programmable data processing device to generate processing implemented by a computer, and instructions executed on the computer or another programmable data processing device provide steps for implementing functions specified in one or more processes in the flowcharts and/or one or more blocks in the block diagrams.

20 In a typical configuration, a computation device includes one or more central processing units (CPUs), input/output interfaces, network interfaces, and memories.

The memory may include the following forms of a computer readable medium: a volatile memory, a random access memory (RAM) and/or a non-volatile memory, for example, a read-only memory (ROM) or flash RAM. The memory is an example of the computer
25 readable medium.

The computer readable medium includes volatile and non-volatile, mobile and non-mobile media, and can use any method or technology to store information. The information may be a computer readable instruction, a data structure, a module of a program or other data. Examples of storage media of the computer include, but are not limited to, a phase
5 change memory (PRAM), a static random access memory (SRAM), a dynamic random access memory (DRAM), other types of RAMs, a ROM, an electrically erasable programmable read-only memory (EEPROM), a flash memory or other memory technologies, a compact disk read-only memory (CD-ROM), a digital versatile disc (DVD) or other optical storage, a cassette tape, a tape disk storage or other magnetic storage devices,
10 or any other non-transmission media, which can be used for storing computer accessible information. According to the definition herein, the computer readable medium does not include transitory computer readable media (transitory media), for example, a modulated data signal and carrier.

It should be further noted that the terms "include," "comprise" or other variations thereof
15 are intended to cover non-exclusive inclusion, so that a process, method, product or device including a series of elements not only includes the elements, but also includes other elements not clearly listed, or further includes inherent elements of the process, method, product or device. In the absence of more limitations, an element defined by "including a/an..." does not exclude that the process, method, product or device including the element
20 further has other identical elements.

As will be appreciated by those skilled in the art, the embodiments of the present application may be embodied as a method, a system, or a computer program product. Accordingly, the present application may use the form of an entirely hardware embodiment, an entirely software embodiment or an embodiment combining software and hardware
25 aspects. Furthermore, the embodiments of the present application may use the form of a

computer program product implemented on one or more computer-usable storage media (including but not limited to magnetic disk memories, CD-ROMs, optical memories, etc.) including computer-usable program code.

The present application may be described in a general context of a computer executable
5 instruction executed by a computer, for example, a program module. Generally, the program module includes a routine, a program, an object, an assembly, a data structure, and the like used for executing a specific task or implementing a specific abstract data type. The present application may also be implemented in distributed computing environments, and in the distributed computer environments, a task is executed by using remote processing devices
10 connected through a communications network. In the distributed computer environment, the program module may be located in local and remote computer storage media including a storage device.

The embodiments in the specification are described progressively, identical or similar parts of the embodiments may be referenced to each other, and each embodiment emphasizes
15 an aspect different from other embodiments. Especially, the system embodiment is basically similar to the method embodiment, and thus is described simply. For related parts, reference may be made to the descriptions of the parts in the method embodiment.

The above descriptions are merely embodiments of the present application, and are not intended to limit the present application. Those skilled in the art can make various
20 modifications and variations to the present application. Any modification, equivalent replacement and improvement made within the spirit and principle of the present application shall fall within the scope of the claims of the present application.

CLAIMS

1. A model training and data similarity determining method, comprising:
 - acquiring a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;
 - acquiring a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair, and acquiring the user similarity corresponding to each user data pair comprises:
 - acquiring biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and
 - determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair;
 - determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model;
 - training the classification model based on the sample data to obtain a similarity classification model;
 - acquiring a to-be-detected user data pair;
 - performing feature extraction on two sets of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features;
 - determining a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and the similarity classification model; and

determining the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair as twins, if the determined similarity is greater than a predetermined similarity classification threshold.

2. The method according to claim 1, wherein the biological features comprise a facial image feature;

the acquiring biological features of users corresponding to a first user data pair comprises:

acquiring facial images of the users corresponding to the first user data pair; and

performing feature extraction on the facial images to obtain facial image features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair comprises:

determining the user similarity corresponding to the first user data pair according to the facial image features of the users corresponding to the first user data pair.

3. The method according to claim 1, wherein the biological features comprise a speech feature;

the acquiring biological features of users corresponding to a first user data pair comprises:

acquiring speech data of the users corresponding to the first user data pair; and

performing feature extraction on the speech data to obtain speech features; and

correspondingly, the determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair comprises:

determining the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

4. The method according to claim 1, wherein the determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a classification model comprises:

performing feature extraction on each user data pair in the plurality of user data pairs to obtain associated user features between the two sets of user data in each user data pair; and

determining, according to the associated user features between the user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

5. The method according to claim 4 wherein the determining, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model comprises:

selecting positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold; and

using the positive sample features and the negative sample features as the sample data for training the classification model.

6. The method according to claim 5 wherein the user feature comprises a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, wherein the household registration dimension feature comprises a feature of user identity information, the name dimension feature comprises a feature of user name information and a

feature of a degree of scarcity of a user surname, and the social feature comprises a feature of social relationship information of a user.

7. The method according to claim 5, wherein the positive sample features comprise the same quantity of features as the negative sample features.

8. The method according to any one of claims 1 to 7, wherein the similarity classification model is a binary classifier model.

9. A model training and data similarity determining apparatus, comprising:
a data acquiring module configured to acquire a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;
a similarity acquiring module configured to acquire a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair, and the similarity acquiring module comprises:
a biological feature acquiring unit configured to acquire biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and
a similarity acquiring unit configured to determine a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair;

a sample data determining module configured to determine, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model;

a model training module configured to train the classification model based on the sample data to obtain a similarity classification model;

a to-be-detected data acquiring module configured to acquire a to-be-detected user data pair;

a feature extraction module configured to perform feature extraction on two sets of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features;

a similarity determining module configured to determine a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and the similarity classification model; and

a similarity classification module configured to determine the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair as twins, if the determined similarity is greater than a predetermined similarity classification threshold.

10. The apparatus according to claim 9, wherein the biological features comprise a facial image feature;

the biological feature acquiring unit is configured to acquire facial images of the users corresponding to the first user data pair and perform feature extraction on the facial images to obtain facial image features; and

correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the facial image features of the users corresponding to the first user data pair.

11. The apparatus according to claim 9, wherein the biological features comprise a speech feature;

the biological feature acquiring unit is configured to acquire speech data of the users corresponding to the first user data pair and perform feature extraction on the speech data to obtain speech features; and

correspondingly, the similarity acquiring unit is configured to determine the user similarity corresponding to the first user data pair according to the speech features of the users corresponding to the first user data pair.

12. The apparatus according to claim 9, wherein the sample data determining module comprises:

a feature extraction unit configured to perform feature extraction on each user data pair in the plurality of user data pairs to obtain associated user features between the two sets of user data in each user data pair; and

a sample data determining unit configured to determine, according to the associated user features between the two sets of user data in each user data pair and the user similarity corresponding to each user data pair, the sample data for training the classification model.

13. The apparatus according to claim 12, wherein the sample data determining unit is configured to select positive sample features and negative sample features from user features corresponding to the plurality of user data pairs according to the user similarity corresponding to each user data pair and a predetermined similarity threshold; and use the positive sample features and the negative sample features as the sample data for training the classification model.

14. The apparatus according to claim 13, wherein the user feature comprises a household registration dimension feature, a name dimension feature, a social feature, and an interest feature, wherein the household registration dimension feature comprises a feature of user identity information, the name dimension feature comprises a feature of user name information and a feature of a degree of scarcity of a user surname, and the social feature comprises a feature of social relationship information of a user.

15. The apparatus according to claim 13, wherein the positive sample features comprise the same quantity of features as the negative sample features.

16. The apparatus according to any one of claims 9 to 15, wherein the similarity classification model is a binary classifier model.

17. A model training and data similarity determining device, comprising:
a processor; and
a memory configured to store computer executable instructions, wherein when being executed, the computer executable instructions cause the processor to execute the following operations:

acquiring a plurality of user data pairs, wherein data fields of two sets of user data in each user data pair have an identical part;

acquiring a user similarity corresponding to each user data pair, wherein the user similarity is a similarity between users corresponding to the two sets of user data in each user data pair, and acquiring the user similarity corresponding to each user data pair comprises:

acquiring biological features of users corresponding to a first user data pair, wherein the first user data pair is any user data pair in the plurality of user data pairs; and

determining a user similarity corresponding to the first user data pair according to the biological features of the users corresponding to the first user data pair;

determining, according to the user similarity corresponding to each user data pair and the plurality of user data pairs, sample data for training a preset classification model;

training the classification model based on the sample data to obtain a similarity classification model;

acquiring a to-be-detected user data pair;

performing feature extraction on two sets of to-be-detected user data in the to-be-detected user data pair to obtain to-be-detected user features;

determining a similarity between users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair according to the to-be-detected user features and the similarity classification model; and

determining the users corresponding to the two sets of to-be-detected user data in the to-be-detected user data pair as twins, if the determined similarity is greater than a predetermined similarity classification threshold.