MULTIMODAL INFORMATION-BASED CLUSTERING METHOD, SERVER, AND STORAGE MEDIUM

[0001] This application claims priority to Chinese Patent Application No. 202310516381.8, filed with the China National Intellectual Property Administration on May 05, 2023 and entitled "MULTIMODAL INFORMATION-BASED CLUSTERING METHOD, SERVER, AND STORAGE MEDIUM", which is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] This application relates to the field of computer technologies, and in particular, to a multimodal information-based clustering method, a server, and a storage medium.

BACKGROUND

[0003] With the development of cloud computing and mobile Internet, service platforms generate massive data. The data is missing multimodality. Different objects include data in different modalities, and different modalities have different attributes. It is very important to eliminate redundancy between modalities by using complementarity between the modalities, to learn better feature attributes. Clustering is one of important applications.

[0004] Currently, mainstream multimodal clustering methods usually perform clustering based on a Euclidean distance or a cosine similarity between features, and the accuracy and recall rate of clustering are low.

SUMMARY

[0005] This application provides a multimodal information-based clustering method, a server, and a storage medium, to resolve problems of low clustering accuracy and low recall rate of the mainstream multimodal clustering methods.

[0006] According to a first aspect, this application provides a multimodal information-based clustering method, including: obtaining multimodal feature information of to-be-clustered target objects; generating a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects; calculating a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality; and clustering the target objects according to the posterior matching probability of any two target objects in multimodality to generate a clustering result of the target objects.

[0007] According to a second aspect, this application provides a multimodal information-based clustering method, including: obtaining multimodal feature information of to-be-clustered images, where the multimodal feature information includes geographic location information and feature information of an included object in at least one modality; generating a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the images; calculating a posterior matching probability of any two images in multimodality according to the posterior matching probability function in each modality, where the posterior matching probability of two images in multimodality is a probability that the two images include a same object; and clustering images including a same object into one category according to the posterior matching probability of any two images in multimodality to obtain a clustering result of the images.

[0008] According to a third aspect, this application provides a multimodal information-based clustering method, applied to a terminal side device, and including: obtaining multimodal feature information of to-be-clustered target objects; sending the multimodal feature information of the to-be-clustered target objects to a server; receiving a clustering result of the target objects sent by the server, where the clustering result of the target objects is generated by using the method according to the first aspect; and executing at least one of the following processing logics according to the clustering result: information pushing, information searching, and data classification.

[0009] According to a fourth aspect, this application provides a server, including a processor and a memory communicatively connected to the processor. The memory stores computer-executable instructions. The processor executes the computer-executable instructions

stored in the memory to implement the method according to the first aspect or the second aspect.

[0010] According to a fifth aspect, this application provides a computer-readable storage medium. The computer-readable storage medium has computer-executable instructions stored therein. The computer-executable instructions, when being executed by a processor, are used for implementing the method according to the first aspect or the second aspect.

BRIEF DESCRIPTION OF DRAWINGS

[0011] Accompanying drawings herein are incorporated into the specification and constitute a part of this specification, show embodiments that conform to this application, and are used for describing a principle of this application together with this specification.

[0012] FIG. 1 is an architectural diagram of an exemplary multimodal information-based clustering system to which this application is applicable.

[0013] FIG. 2 is a flowchart of a multimodal information-based clustering method according to an exemplary embodiment of this application.

[0014] FIG. 3 is a flowchart of a multimodal information-based clustering method according to another exemplary embodiment of this application.

[0015] FIG. 4 is a framework diagram of determining a posterior matching probability of target objects in multimodality based on multimodal information according to an exemplary embodiment of this application.

[0016] FIG. 5 is a flowchart of a multimodal information-based clustering method according to another exemplary embodiment of this application.

[0017] FIG. 6 is an interaction diagram of a multimodal information-based clustering method according to an exemplary embodiment of this application.

[0018] FIG. 7 is a schematic structural diagram of a multimodal information-based clustering apparatus according to an exemplary embodiment of this application.

[0019] FIG. 8 is a schematic structural diagram of a server according to an embodiment of this application.

[0020] The foregoing accompanying drawings have shown explicit embodiments of this application, and more detailed descriptions are provided below. These accompanying drawings

and text descriptions are not intended to limit the scope of concept of this application, but to describe the concept of this application for a person skilled in the art with reference to specific embodiments.

DESCRIPTION OF EMBODIMENTS

[0021] Exemplary embodiments are described in detail herein, and examples of the exemplary embodiments are shown in the accompanying drawings. When the following description involves the accompanying drawings, unless otherwise indicated, the same numerals in different accompanying drawings represent the same or similar elements. The implementations described in the following exemplary embodiments do not represent all implementations that are consistent with this application. On the contrary, the implementations are merely examples of apparatuses and methods that are described in detail in the appended claims and that are consistent with some aspects of this application.

[0022] Terms in this application are first interpreted.

[0023] Modality: Each type of source or form of information may be referred to as a modality. For example, people have touch, auditory, visual, olfactory, and gustatory senses, which may be used as a plurality of different modalities. Information media include voice, video, text, and the like, and different media may be used as a plurality of different modalities. There are various sensors, such as radar, infrared, and accelerometer, and different source devices may be used as different modalities. For example, multimodal feature information of a person may include visual information such as various biological features, basic profile information, and geographical information such as time and space. In an e-commerce scenario, multimodal feature information of an item may include an image feature, a text description, an item category, sales volume information, comment information, and the like.

[0024] Clustering: A process in which a set of physical or abstract objects is divided into a plurality of clusters formed by similar objects is referred to as clustering. A cluster generated through clustering is a set of a group of objects. These objects and objects in the same cluster are similar to each other, and are different from objects in another cluster.

[0025] Multimodal clustering: Clustering objects having multimodal information. In an

example scenario, images of a same object among images acquired in a plurality of different regions are aggregated into a same category based on multimodal information of the clustered objects. In an e-commerce scenario, items in different shops are aggregated into a same category according to an item uniqueness dimension based on multimodal information of the items.

[0026] Missing multimodal: Information in one or more modalities is missing for some objects in an object set having multimodal information. For example, some objects do not have associated persons, and some items do not have associated comment information.

[0027] Euclidean distance: A distance between two points in Euclidean space.

[0028] Cosine similarity: A similarity between two vectors is measured by calculating a cosine value of an included angle of the two vectors.

[0029] Matching probability: A probability that two objects belong to a same clustering category (cluster).

[0030] Conditionally independent: If $P(X, Y|Z) = P(X|Z)P(Y|Z)$, or equivalently, $P(X|Y, Z) = P(X|Z)$, it means that events X and Y are conditionally independent for a given event Z. In other words, when Z occurs, whether X occurs is irrelevant to whether Y occurs. In this embodiment, multiple modalities need to be conditionally independent for a given category label and a defined scenario. For example, similarities of different biological feature are conditionally independent given whether it is a same person, and a biological feature similarity and a spatial distance are conditionally independent given whether it is a same person.

[0031] Posterior matching probability in a single modality: A posterior matching probability determined based on single-modal feature information of objects.

[0032] Posterior matching probability in multimodality: A posterior matching probability between objects determined based on multimodal feature information of the objects.

[0033] Posterior matching probability function in a single modality: A posterior matching probability function based on single-modal feature information.

[0034] Posterior matching probability function in multimodality: A posterior matching probability function based on multimodal feature information of objects, including a cross posterior matching probability function in two modalities and a mixed posterior matching probability function of three or more modalities.

[0035] With the development of cloud computing and the mobile Internet, service platforms

generate massive data. For example, large-scale short video data that is photographed and uploaded by mobile terminals, and online clustering is performed by obtaining multimodal information such as video content or images, sounds, and texts, which may be applied to application scenarios such as clustering and retrieval of videos of a same type. For another example, expression manners of people on social networks become more and more abundant. People may express their emotions and opinions by using pictures, texts, videos, and the like. Multimodal information such as pictures, videos, and texts is utilized for dynamic content clustering, dynamic emotion clustering, and the like on social networks to recommend similar dynamic content to users. For another example, an increasing quantity of shopping platforms can provide search services. Different merchants in the shopping platforms have multimodal information such as different descriptions of videos, pictures, texts, prices and the like for items of a same type. Through multimodal information clustering, similar items of from different merchants are clustered together, and a search service combining videos, pictures, and texts can be provided for the users.

[0036] Currently, mainstream multimodal clustering methods usually perform clustering based on a Euclidean distance or a cosine similarity between features, and accuracy and a recall rate of clustering are low. In some multimodal clustering methods, a Euclidean distance or a cosine similarity of a key feature is used as a similarity between objects, and feature information in another modality is used as a rule for optimizing the similarity between the objects. For example, when capturing addresses of two pictures are located in a same street, the similarity is increased by 5%; when capturing addresses of two pictures are located in a same region, the similarity is increased by 2%; or the like. Rule setting is not flexible enough, and a clustering effect is affected.

[0037] To resolve the problem that the accuracy and recall rate of clustering are low in the current mainstream multimodal clustering methods, this application provides a multimodal information-based clustering method. In this method, a posterior matching probability function in each single modality is generated through self-learning according to multimodal feature information of to-be-clustered target objects, so that the posterior matching probability function in each modality can be automatically learned without using annotation information; a posterior matching probability of any two target objects in multimodality is calculated according to the

posterior matching probability function in each single modality; and the target objects are clustered according to the posterior matching probability of any two target objects in multimodality to generate a clustering result of the target objects. The posterior matching probability that the target objects belong to a same cluster is calculated based on the multimodal feature information of the target objects by using the posterior matching probability function generated through self-learning, and is used as a metric distance for the similarity between the target objects. In comparison with using the Euclidean distance or the cosine similarity between features in the mainstream methods, using the posterior matching probability between the target objects can measure the similarity between the target objects more accurately. Clustering is performed based on the posterior matching probability between the target objects, which can greatly improve the accuracy and recall rate of the clustering.

[0038] The multimodal information-based clustering method provided in this application can generate the posterior matching probability function in each single modality through self-learning, and thereby can be quickly applied to different data distribution scenarios in an unbiased manner, for example, aggregation and collation of picture libraries of different regions and clustering of similar commodities.

[0039] FIG. 1 is an architectural diagram of an exemplary multimodal information-based clustering system to which this application is applicable. As shown in FIG. 1, the system architecture may specifically include a server, a terminal side device, and a data service device.

[0040] The server may be specifically a server locally deployed, or may be a server cluster deployed in a cloud. There is a communicable communication link between the server and each terminal side device, enabling communication connection between the server and each terminal side device.

[0041] The terminal side device refers to an electronic device providing multimodal feature information of to-be-clustered target objects to the server, and may be specifically a hardware device having a network communication function, a computing function, and an information display function, including but not limited to, a terminal device such as an intelligent terminal, a tablet computer, or a personal computer used by a user, an Internet of Things device, a server of a service platform or an organization, and the like.

[0042] The data service device may be a device that can provide multimodal information

for the server. The data service device may be specifically a data collection device, or a server that can collect data in one or more modalities collected by a data collection device. The server may obtain multimodal information of the target objects from one or more data service devices, and perform data preprocessing and feature extraction, to obtain the multimodal feature information of the target objects.

[0043] The terminal side device sends a clustering request/instruction for the target objects to the server, or invokes an application programming interface provided by the server, to trigger the server to perform clustering based on the multimodal information of the target objects. In response to the clustering request/instruction for the target object, or the application programming interface being invoked, the server obtains the multimodal information of the to-be-clustered target objects from the one or more data service devices, and generates multimodal feature information of the target objects. The server generates a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects; calculates a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality; and clusters the target objects according to the posterior matching probability of any two target objects in multimodality, to generate a clustering result of the target objects. The server sends the clustering result to the terminal side device, and the terminal side device may execute at least one of the following processing logics according to the clustering result: information pushing, information searching, and data classification. In another implementation, the server may execute at least one of the following processing logics according to the clustering result: information pushing, information searching, and data classification, and send an execution result to the terminal side device.

[0044] In another implementation, the terminal side device may obtain multimodal data of the target objects from the data service device. Data in different modalities may be from different devices. The multimodal data of the target objects is preprocessed to generate the multimodal feature information of the target objects, and the multimodal feature information of the target objects is sent to the server.

[0045] A possible application scenario is a scenario of clustering and collating image libraries of different regions. Images that belong to a same object in the images captured in

different regions are clustered to a same cluster based on geographic location information of image capturing and feature information that is in at least one modality and of an object included in the images. For example, based on images captured in different regions, multimodal clustering is performed on the images based on geographic location information of image capturing, and attributive features and biometric features of persons included in the images, to aggregate images of a same person to a same cluster. For example, based on motor vehicle images captured in different regions, multimodal clustering is performed on the motor vehicle images based on geographic location information of motor vehicle image capturing, license plate numbers, owner identities, and appearance features of motor vehicles included in the motor vehicle images, and the like, to aggregate images of a same motor vehicle into a same cluster.

[0046] In this scenario, the to-be-clustered target objects are images, and the server may be a server of a system platform that has an information library aggregation and collation requirement and has authorization, and may be specifically a server deployed on a local or cloud platform. The server obtains to-be-clustered images and geographic location information of the images from inside the system platform or from an image capturing device for which the system platform has data authorization, recognizes objects included in the images by using an image recognition technology, and obtains feature information in at least one modality of the included objects, to obtain multimodal feature information of the to-be-clustered images. The server generates a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the images; calculates a posterior matching probability of any two images in multimodality according to the posterior matching probability function in each modality, where the posterior matching probability of the two images in multimodality is a probability that the two images include a same object; and clusters images including a same object into one category according to the posterior matching probability of any two images in multimodality to obtain a clustering result of the images. For example, according to a clustering result, in accordance with clusters to which the images belong, the server associates images of a same cluster with an included object and stores the images by classification to implement the function of information library aggregation and collation. The to-be-clustered information in the information library may be images, audios, or another type

of data information, which is not specifically limited herein.

[0047] Another possible application scenario is an e-commerce scenario. Different merchants on an e-commerce platform have information in different modalities, such as videos, pictures, introduction texts, and price information, for a same type of items. Similar items from different merchants are clustered to a same cluster through multimodal clustering, to implement services such as item searching and item recommendation based on the multimodal information such as videos, pictures, introduction texts, and price information.

[0048] In this scenario, the to-be-clustered target objects are items, and the server may be a server of the e-commerce platform, and may be specifically a server deployed on a local or cloud platform. The server obtains, from within the e-commerce platform, the multimodal information such as videos, pictures, introduction texts, and price information of different merchants for the items, and performs data preprocessing to generate multimodal feature information of the items. The server generates a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the items; calculates a posterior matching probability of any two items in multimodality according to the posterior matching probability function in each modality, where the posterior matching probability of the two items in multimodality is a probability that the two items are of a same type; and clusters items of a same type into one type according to the posterior matching probability of any two items in multimodality to obtain a clustering result. For example, based on the clustering result, the server may associate items that belong to a same cluster, search for similar items according to query information inputted by a user or an item in which the user is currently interested (for example, clicks to browse, adds to cart, or follows), and recommend the similar items to the user.

[0049] The method of this embodiment may be specifically applied to scenarios in which clustering needs to be performed in various application fields, including but not limited to fields such as information library aggregation and collation, e-commerce, education, and medical treatment. Abstract objects, such as data objects, and physical objects may be clustered. Types of to-be-clustered target objects and specific application scenarios of clustering results are not specifically limited herein.

[0050] The following describes the technical solutions of this application and how the

technical solutions of this application resolve the foregoing technical problems in detail by using specific embodiments. The following specific embodiments may be combined with each other, and descriptions of the same or similar concepts or processes may not be repeated in some embodiments. The embodiments of this application will be described below with reference to the accompanying drawings.

[0051] FIG. 2 is a flowchart of a multimodal information-based clustering method according to an exemplary embodiment of this application. An executive entity of this embodiment may be the server in the multimodal information-based clustering system architecture mentioned above. The method provided in this embodiment is used for implementing multimodal clustering of target objects based on multimodal feature information of the target objects. As shown in FIG. 2, the method specifically includes the following steps.

[0052] Step S201: Obtain multimodal feature information of to-be-clustered target objects.

[0053] In this embodiment, to accurately cluster a plurality of to-be-clustered target objects, the server obtains feature information of the target objects in a plurality of different modalities. In different application scenarios, sources and manners for the server to obtain the multimodal feature information may be different.

[0054] For example, the server may obtain the multimodal feature information of the target objects from one or more data service devices inside a system platform where the server is located. Feature information in different modalities may be from a same data service device or different data service devices.

[0055] For example, the server may also obtain multimodal data of the target objects from one or more data service devices inside a system platform where the server is located, and perform data preprocessing and feature extraction to generate the multimodal feature information of the target objects.

[0056] For example, the server may also receive the multimodal feature information of the target objects sent by a terminal side device. The terminal side device may be a device that sends a clustering request/instruction to the server, or invokes a clustering application programming interface provided by the server. The terminal side device may perform data preprocessing and feature extraction on the multimodal data of the target objects provided by one or more data service devices, to obtain the multimodal feature information of the target

objects.

[0057] The target object is any one of the following: a physical object, an image including a physical object, or a text including a physical object. The multimodal feature information of the target object includes at least one of the following: an image feature, a text feature, an attribute feature, or a geographic location feature.

[0058] The multimodal feature information of the target object includes, but is not limited to: attribute information of the target object in one or more modalities (which may include a geographic location feature, a category feature, and the like), a feature of content included by the target object in one or more modalities (which may include an image feature, a text feature, and the like), and a feature of an object associated with the target object in one or more modalities (which may include an image feature, a text feature, an attribute feature, a geographic location feature, and the like).

[0059] For example, in an image library aggregation and collation scenario, to-be-clustered target objects may be images, and multimodal feature information of the target objects includes, but is not limited to: geographic location information of image capturing, feature information of an object included in the images in at least one modality, such as a biological feature of a person included in the images, and a license plate number, owner information, and an appearance feature of a motor vehicle included in the images.

[0060] For example, in an e-commerce scenario, to-be-clustered target objects may be items, and multimodal feature information of the target objects includes, but is not limited to, videos, images, introduction texts, and price information.

[0061] Step S202: Generate a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects.

[0062] In this embodiment, the posterior matching probability function in each modality is automatically generated based on the multimodal feature information of the target objects through self-learning. The posterior matching probability function in any modality may be used for calculating, based on feature information of the target objects in the modality, a posterior matching probability that any two target objects belong to a same cluster. The posterior matching probability reflects a similarity between the target objects in the feature information in the current modality.

[0063] Step S203: Calculate a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality.

[0064] After the posterior matching probability function in each modality is generated through self-learning, the posterior matching probability of any two target objects in multimodality is calculated and determined based on the posterior matching probability function in each modality and the multimodal feature information of the target objects. This posterior matching probability reflects an overall similarity between the multimodal feature information of the target objects. In comparison with a posterior matching probability calculated based on feature information in a single modality, using the posterior matching probability of different target objects in multimodality is more accurate in measuring the similarity between the target objects.

[0065] Step S204: Cluster the target objects according to the posterior matching probability of any two target objects in multimodality to generate a clustering result of the target objects.

[0066] After the posterior matching probability of any two target objects in multimodality is obtained, the target objects are clustered based on the posterior matching probability of the any two target objects in multimodality, target objects having a larger posterior matching probability in multimodality are clustered into a same cluster, and target objects having a smaller posterior matching probability in multimodality are clustered into different clusters, thereby implementing accurate clustering of the target objects.

[0067] In this step, any existing clustering algorithm may be used, and the posterior matching probability of the target objects in multimodality is used as a metric distance for the similarity between the target objects to perform clustering. A specific clustering procedure is not repeated here.

[0068] According to the method in this embodiment, a posterior matching probability function in each single modality is generated through self-learning according to multimodal feature information of to-be-clustered target objects, so that the posterior matching probability function in each modality can be automatically learned without using annotation information; a posterior matching probability of any two target objects in multimodality is calculated according to the posterior matching probability function in each single modality; and the target objects are clustered according to the posterior matching probability of any two target objects

in multimodality to generate a clustering result of the target objects. The posterior matching probability that the target objects belong to a same cluster is calculated based on the multimodal feature information of the target objects by using the posterior matching probability function generated through self-learning, and is used as a metric distance for the similarity between the target objects. In comparison with using the Euclidean distance or the cosine similarity between features in mainstream methods, using the posterior matching probability between the target objects can measure the similarity between the target objects more accurately. Clustering is performed based on the posterior matching probability between the target objects, which can greatly improve the accuracy and recall rate of the clustering.

[0069] FIG. 3 is a flowchart of a multimodal information-based clustering method according to another exemplary embodiment of this application. Based on the foregoing method embodiments, the multimodal information-based clustering method is described in detail in this embodiment. As shown in FIG. 3, specific steps of the method are as follows.

[0070] Step S301: Obtain multimodal feature information of to-be-clustered target objects.

[0071] This step is implemented in the same manner as the foregoing step S201, and details are not described here again.

[0072] In an optional embodiment, a server may receive the multimodal feature information of the to-be-clustered target objects sent by a terminal side device. Further, after obtaining a final clustering result of the target objects, the server outputs the clustering result of the target objects to the terminal side device.

[0073] After the multimodal feature information of the to-be-clustered target objects is obtained, step S302 to S304 are performed to implement step S202 of generating a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects.

[0074] Step S302: Calculate a similarity between any two target objects in a same modality according to feature information of the target objects in the same modality.

[0075] Specifically, the feature information of the target objects in each modality may be encoded into a feature vector. Based on feature vectors in any modality, a Euclidean distance or a cosine similarity between feature vectors of any two target objects in the modality is calculated, and is used as a similarity between the any two target objects in the modality.

[0076] The similarity between any two target objects in any modality may be stored as a matrix to obtain a similarity matrix in the modality. M is used to represent a quantity of modalities, and similarity matrices in the modalities may be represented as $w^{(1)}$, $w^{(2)}$, …, and $w^{(M)}$.

[0077] Step S303: Construct a plurality of sample pairs in each modality according to the similarity between any two target objects in each modality, where the sample pair includes two target objects.

[0078] In this step, for any modality, a plurality of sample pairs in the modality are constructed based on the similarity between any two target objects in the modality. Sample pairs in each modality are constructed based on the same manner to obtain a plurality of sample pairs in each modality.

[0079] In a possible implementation of this step, when a plurality of sample pairs in any modality are constructed, any two target objects may form a sample pair; or two different target objects may be randomly selected to form a sample pair.

[0080] In another possible implementation of this step, when a plurality of sample pairs in any modality are constructed, K nearest neighbor objects of each target object in the modality may also be determined according to the similarity between any two target objects in the modality; and any target object separately forms a sample pair with each of the K nearest neighbor objects in this modality to obtain K sample pairs including the target object. A quantity of the target objects is represented as N, and N×K sample pairs in this modality can be obtained. K and N are both positive integers. By constructing high-quality sample pairs in this manner, the quantity of the sample pairs is reduced, and quality and efficiency of generating the posterior matching probability function in each modality through self-learning can be improved.

[0081] The K nearest neighbor objects of any target object in the modality are obtained by determining, based on similarities between other target objects and the target object in the modality, K target objects having high similarities with the target object by using a K nearest neighbor (K Nearest Neighbor, KNN for short) algorithm. K may be set and adjusted according to a requirement of an actual application scenario, which is not specifically limited herein.

[0082] Step S304: Generate a posterior matching probability function in each modality through self-learning by using an optimization algorithm and according to similarities of the

plurality of sample pairs in each modality.

[0083] After the plurality of sample pairs in each modality are constructed, a single-modal posterior matching probability function in each modality may be generated through self-learning by using an optimization algorithm and based on the constructed sample pairs and the similarities of the sample pairs. Specifically, this may be implemented in the following steps S3041 to S3044:

[0084] Step S3041: Initialize the posterior matching probability function in each modality to a monotonically increasing piecewise function, where the piecewise function includes to-be-solved segment parameters.

[0085] In an actual application, as the similarity between the target objects increases, the posterior matching probability between the target objects also increases. In this embodiment, the sample pairs in each modality are sorted according to an ascending order of similarities, and then the sample pairs are divided into I segment sets, where I is a positive integer. T is used to represent a quantity of sample pairs in a single modality, T = N×K, and a quantity of sample pairs included in each segment set is $\frac{T}{I}$. The single-modal posterior matching probability function in each modality is initialized to a monotonically increasing piecewise function including I segments. M is used to represent a quantity of modalities, and a posterior matching probability function in an mth modality may be represented as the following piecewise function:

$$f^{(m)}(x): (x_{i_m}^{(m)}, f_{i_m}^{(m)})$$

$$\text{subject to (s.t.):} \quad x_1^{(m)} < x_2^{(m)} < \ldots < x_I^{(m)},$$

$$f_1^{(m)} < f_2^{(m)} < \ldots < f_I^{(m)}, \tag{1}$$

$$f_1^{(m)} = 0,\ f_I^{(m)} = 1,$$

$$|z_{i_m}^{(m)}| = length\left(r_{i_m}^{(m)}\right) = \frac{T}{I}$$

[0086] In the formula (1), $z_{i_m}^{(m)}$ represents an i_mth segment set in the mth modality. $|z_{i_m}^{(m)}|$ represents a quantity of sample pairs included in the i_mth segment set in the mth modality. $r_{i_m}^{(m)}$ represents a set of similarities of sample pairs in the i_mth segment set $z_{i_m}^{(m)}$ in the mth modality.

$length\left(r_{i_m}^{(m)}\right)$ represents a quantity of similarity values in $r_{i_m}^{(m)}$. $x_{i_m}^{(m)}$ represents an average value of the similarities of sample pairs in the i_mth segment set in the mth modality. Because the segment sets are obtained by dividing the sample pairs sorted in the ascending order of similarities, $x_1^{(m)} < x_2^{(m)} < \ldots < x_I^{(m)}$ holds. $f_{i_m}^{(m)}$ represents a function value of an i_mth segment of the posterior matching probability function in the mth modality, and is a to-be-solved segment parameter. Since the piecewise function is a monotonically increasing function, $f_1^{(m)} < f_2^{(m)} < \ldots < f_I^{(m)}$ holds. The minimum value of the posterior matching probability is 0, and the maximum value is 1, so $f_1^{(m)} = 0$ and $f_I^{(m)} = 1$ holds. $i_m \in$ [1, 2, …, I], m $\in$ [1, 2, …, M].

[0087] The posterior matching probability function in each modality can be initialized to the monotonically increasing piecewise function by using the formula (1). An independent variable of each segment is an average value of similarities of sample pairs in a corresponding segment set, the function value is a to-be-solved segment parameter, and there are M×I to-be-solved segment parameters in total.

[0088] It should be noted that, the quantity I of segments of the piecewise function may be set according to an actual application scenario. Different values may be used in different application scenarios, and this is not specifically limited herein. For example, a value of I may be 1000, 200, or the like.

[0089] Step S3042: Construct an objective function based on the posterior matching probability function in each modality, where the objective function includes the to-be-solved segment parameters.

[0090] In this embodiment, the posterior matching probability function in each modality is the single-modal posterior matching probability function, and a single-modal posterior matching probability function in the mth modality may be represented as $Fsingle_{i_m}^{(m)}$:

$$Fsingle_{i_m}^{(m)} = f_{i_m}^{(m)} = f^{(m)}\left(x_{i_m}^{(m)}\right) \quad (2)$$

[0091] In this step, a posterior matching probability function in multimodality is constructed according to the single-modal posterior matching probability function in each

modality.

[0092] Specifically, a cross posterior matching probability function of any two modalities may be constructed. A cross posterior matching probability function in the m^{th} modality and a b^{th} modality may be represented as $Fcross_{i_m}^{(m)-(b)}$:

$$Fcross_{i_m}^{(m)-(b)} = \frac{1}{|z_{i_m}^{(m)}|}\sum_{x\in r_{i_m}^{(m)},\, x_{i_b}^{(b)}\in r_{i_m}^{(m)-(b)}} f^{(b)}\left(x_{i_b}^{(b)}|x\right) = \frac{1}{|z_{i_m}^{(m)}|}\sum_{x\in r_{i_m}^{(m)},\, x_{i_b}^{(b)}\in r_{i_m}^{(m)-(b)}} f_{i_b}^{(b)} \quad (3)$$

[0093] In the formula (3), $r_{i_m}^{(m)-(b)}$ represents a set of average values of similarities of sample pairs in segment sets in the b^{th} modality to which the sample pairs in the $i_m{}^{th}$ segment set in the m^{th} modality belong. $x_{i_b}^{(b)}$ is an average value of similarities in $r_{i_m}^{(m)-(b)}$. $i_m, i_b \in [1, 2, \ldots, I]$. $m, b \in [1, 2, \ldots, M]$. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0094] Further, for similarities $\{x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_M}^{(M)}\}$ between feature information of any two target objects in given M modalities, a joint probability distribution function of the M modalities may be represented as:

$$Fjoint\left(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_M}^{(M)}\right) = \frac{\prod_{m=1}^{M} f^{(m)}\left(x_{i_m}^{(m)}\right)}{\prod_{m=1}^{M} f^{(m)}\left(x_{i_m}^{(m)}\right) + \prod_{m=1}^{M}\left(1 - f^{(m)}\left(x_{i_m}^{(m)}\right)\right)} = \frac{\prod_{m=1}^{M} f_{i_m}^{(m)}}{\prod_{m=1}^{M} f_{i_m}^{(m)} + \prod_{m=1}^{M}\left(1-f_{i_m}^{(m)}\right)} \quad (4)$$

[0095] Further, a posterior matching probability function of the M modalities is constructed, and may be represented as:

$$Fmulti_{i_m}^{(m)} = \frac{1}{|z_{i_m}^{(m)}|}\sum_{x\in r_{i_m}^{(m)}} Fjoint\left(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_M}^{(M)}|x\right) \quad (5)$$

[0096] Meanings of characters in the formulas (4) and (5) are the same as those in the foregoing content, and details are not described herein again.

[0097] In an actual application, for any set of sample pairs, an average value of probabilities that sample pairs belong to a same category as statistically derived based on one modality should be equal to an average value of probabilities that the sample pairs belong to a same category as statistically derived based on any plurality of modalities, an expected value of which

should be equal to a true proportion of sample degrees belonging to a same category. In other words, for a same set of sample pairs, posterior matching probability distributions determined based on a posterior matching probability function of a single modality, based on a cross posterior matching probability function of two modalities, and based on a posterior matching probability function of three or more modalities should be consistent. Based on this, the consistency loss function is constructed according to the posterior matching probability function in each modality and the posterior matching probability function in multimodality.

[0098] For example, the following consistency loss function may be constructed:

$$L_{consistency} = \frac{1}{M}\sum_m \left(\frac{1}{I}\sum_{i_m} D\left(Fsingle_{i_m}^{(m)}, Fmulti_{i_m}^{(m)}\right)\right) + \frac{1}{M}\sum_m \left(\frac{1}{I}\sum_{i_m}\sum_{b\neq m} D\left(Fsingle_{i_m}^{(m)}, Fcross_{i_m}^{(m)-(b)}\right)\right) \quad (6)$$

[0099] D( ) represents a distance function, which can obtain a Euclidean distance or a cosine similarity. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0100] In the formula (6), because $Fsingle_{i_m}^{(m)}$ is constrained by $Fmulti_{i_m}^{(m)}$ and $Fcross_{i_m}^{(m)-(b)}$, it is difficult to learn a consistency loss of the three in the formula (6). Based on $Fsingle_{i_m}^{(m)}$, $Fmulti_{i_m}^{(m)}$, and $Fcross_{i_m}^{(m)-(b)}$, a mixed posterior matching probability function is defined as follows:

$$Fmix_{i_m}^{(m)} = \sqrt{Fmulti_{i_m}^{(m)} \frac{1}{M}\left(Fsingle_{i_m}^{(m)} + \sum_{b\neq m} Fcross_{i_m}^{(m)-(b)}\right)} \quad (7)$$

[0101] The mixed posterior matching probability function in the formula (7) is used to constrain values of $Fsingle_{i_m}^{(m)}$ and $Fcross_{i_m}^{(m)-(b)}$. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0102] Further, based on the formula (7), the consistency loss function in the formula (6) may be replaced with a consistency loss function in the following formula (8):

$$L_{consistency} = \frac{1}{M}(L_{consistency} + L_{consistency2}) \quad (8)$$

where $L_{consistency1} = \frac{1}{M}\sum_m \left(\frac{1}{I}\sum_{i_m} D\left(Fsingle_{i_m}^{(m)}, Fmix_{i_m}^{(m)}\right)\right)$, and $L_{consistency2} = \frac{1}{M}\sum_m \left(\frac{1}{I}\sum_{i_m}\sum_{b\neq m} D\left(Fcross_{i_m}^{(m)-(b)}, Fmix_{i_m}^{(m)}\right)\right)$. $L_{consistency1}$ constrains consistency

between the single-modal posterior matching probability function and the mixed posterior matching probability function; and $L_{consistency2}$ constrains consistency between the cross posterior matching probability function of two modalities and the mixed posterior matching probability function.

[0103] In an actual application, regardless of the posterior matching probability function of a single modality, the cross posterior matching probability function of two modalities, or the mixed posterior matching probability function, the minimum value thereof is 0, and the maximum value thereof is 1. Based on this, a constraint loss function is constructed.

[0104] For example, the following constraint loss function may be constructed:

$$\begin{aligned} L_{constraint} = & \sum_m \left(\sum_{i \in r_i} D\left(Fmulti_i^{(m)}, 0\right) + \sum_{j \in r_j} D\left(Fmulti_j^{(m)}, 1\right) \right) \\ & + \sum_m \left(\sum_{i \in r_i} D\left(Fsingle_i^{(m)}, 0\right) + \sum_{j \in r_j} D\left(Fsingle_j^{(m)}, 1\right) \right) \\ & + \sum_m \left(\sum_{b \neq m} \left(\sum_{i \in r_i} D\left(Fcross_i^{(m)-(b)}, 0\right) + \sum_{j \in r_j} D\left(Fcross_i^{(m)-(b)}, 1\right) \right) \right) \end{aligned} \quad (9)$$

where $r_i \in [1,2,\dots,indi]$ indicates the first $indi$ segments of the piecewise function; and $r_j \in [I-indj,\dots,I-1,I]$ indicates the last $indj$+1 segments of the piecewise function. Values of $indi$ and $indj$ may be set and adjusted according to an actual application scenario, and the values are different in different application scenarios. This is not specifically limited herein. For example, in some scenarios, $indi = 10$, and $indj = 4$. In some other scenarios, $indi = 10$, and $indj = 2$.

[0105] Further, according to the consistency loss function in the formula (8) and the constraint loss function in the formula (9), the following objective function is constructed:

$$L = \lambda L_{consistency} + L_{constraint} \quad (10)$$

[0106] In the formula (10), λ is a preset weight parameter, used for balancing the consistency loss and the constraint loss. A value of λ may be set according to an empirical value, and is not specifically limited herein. The value may be different in different application scenarios, for example, may be 80, 20, 2, etc. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0107] In an optional embodiment, the objective function may be constructed based on only the foregoing constraint loss function.

[0108] Step S3043: Minimize the objective function by using the optimization algorithm

according to the similarities of the plurality of sample pairs in each modality, to solve and obtain the segment parameters.

[0109] After the objective function is constructed, segment parameters enabling the objective function to have a minimum value is solved by using an optimization algorithm and according to the similarities of the plurality of sample pairs in each modality, which may be specifically implemented by using any optimization algorithm among a gradient descent method, a Newton method, a quasi-Newton method, a momentum optimization method, a conjugate gradient method, and a heuristic optimization algorithm. This is not specifically limited herein.

[0110] Step S3044: Substitute the segment parameters into the piecewise function to obtain the posterior matching probability function in each modality.

[0111] After the segment parameters are obtained through solving, the segment parameters are substituted into the piecewise function to obtain the posterior matching probability function in each modality.

[0112] In another optional implementation of step S304, the posterior matching probability function in each modality may alternatively be initialized to a continuous monotonous function, where the continuous monotonous function includes a to-be-solved unknown parameter. A consistency loss function and a constraint loss function are constructed based on the continuous monotonous function, and the objective function is constructed. The unknown parameter is obtained through solving by minimizing the objective function by using an optimization algorithm, to obtain a continuous monotonous posterior matching probability function in each modality.

[0113] After the posterior matching probability function in each modality is learned, step S203 is implemented by using steps S305 and S306 to calculate a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality.

[0114] Step S305: Determine, according to the similarity between any two target objects in any modality, a posterior matching probability of the any two target objects in the modality by using the posterior matching probability function in the modality.

[0115] After the posterior matching probability function in each modality is obtained, for

any modality, using the m[th] modality as an example, according to a similarity between any two target objects in the modality, in independent variables $\{x_1^{(m)}, x_2^{(m)}, \ldots, x_I^{(m)}\}$ of I segments of the posterior matching probability function in the modality, an independent variable having the minimum absolute difference with the similarity of the two target objects is determined, and a posterior matching probability value corresponding to the independent variable is used as a posterior matching probability of the two target objects in the modality. Similarly, the posterior matching probability of any two target objects in each modality can be determined.

[0116] Step S306: Calculate a posterior matching probability of any two target objects in multimodality according to the posterior matching probability of the any two target objects in each modality.

[0117] In this step, after the posterior matching probability of any two target objects in each modality is calculated, a posterior matching probability $P(i,j)$ of the any two target objects i and j in multimodality may be calculated by using the following formula (11):

$$P(i,j) = \frac{\prod_{m=1}^{M} P(e_{ij}=1|w_{ij}^{(m)})}{\prod_{m=1}^{M} P(e_{ij}=1|w_{ij}^{(m)}) + \prod_{m=1}^{M} P(e_{ij}=0|w_{ij}^{(m)})} \quad (11)$$

where $P(e_{ij}=1|w_{ij}^{(m)})$ represents a posterior matching probability of the target objects i and j in the m[th] modality, and $P(e_{ij}=0|w_{ij}^{(m)}) = 1 - P(e_{ij}=1|w_{ij}^{(m)})$. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0118] Optionally, in this step, an average value of posterior matching probabilities of the any two target objects in the modalities may be calculated as the posterior matching probability of the two target objects in multimodality. Optionally, in this step, posterior matching probabilities of any two target objects in the modalities may be weighted to obtain an average value according to weight coefficients corresponding to the modalities, and a calculation result is used as the posterior matching probability of the two target objects in multimodality.

[0119] For example, FIG. 4 is a framework diagram of determining a posterior matching probability of target objects in multimodality based on multimodal information according to this embodiment. As shown in FIG. 4, feature information of the target objects in M modalities is first obtained. For any modality, a similarity between feature information of any two target

objects in the modality is calculated, to obtain similarity matrices $w^{(1)}$, …, and $w^{(M)}$ of the modalities. According to the similarity matrices $w^{(1)}$, …, and $w^{(M)}$ of the modalities, T sample pairs (p_1, q_1), (p_2, q_2), …, and (p_T, q_T) in each modality are constructed by using a KNN algorithm, and similarities of the T sample pairs are determined, for example, in the mth modality, a similarity of (p_1, q_1) may be represented in a form of $w^{(1)}_{p_1,q_1}$. The T sample pairs in each modality are sorted according to similarities, and then divided into I segment sets, where a_x, b_y, …, and c_z respectively represent a quantity of sample pairs in the I segment sets; $w^{(1)}_{a_1}$, $w^{(1)}_{a_2}$, …, and $w^{(1)}_{a_x}$ respectively represent similarities of a_x sample pairs in the first segment set in the first modality; $w^{(1)}_{b_1}$, $w^{(1)}_{b_2}$, …, and $w^{(1)}_{b_y}$ respectively represent similarities of b_y sample pairs in the second segment set in the first modality; and $w^{(1)}_{c_1}$, $w^{(1)}_{c_2}$, …, and $w^{(1)}_{c_z}$ respectively represent similarities of c_z sample pairs in the last segment set in the first modality. Representation forms of similarities of sample pairs in segment sets in another modality are similar thereto, for example, $w^{(2)}_{a_1}$, $w^{(2)}_{a_2}$, …, and $w^{(2)}_{a_x}$ respectively represent similarities of a_x sample pairs in the first segment set in the second modality. This is not listed one by one herein. The posterior matching probability function in each single modality is initialized to a piecewise function, a cross posterior matching probability function of two modalities (that is, cross-modality) and a posterior matching probability function of the M modalities are constructed based on the piecewise functions, and the posterior matching probability function in each modality is generated through self-learning based on consistency of various types of posterior matching probability functions. Based on the posterior matching probability function in each modality and a similarity between any two target objects in each modality, a posterior matching probability of the any two target objects in each single modality may be calculated. Further, based on the posterior matching probability of the any two target objects in each single modality, a posterior matching probability of the any two target objects in multimodality may be calculated, as a final posterior matching probability of the two target objects in multimodality. Further, a clustering result of the target objects may be obtained by performing clustering based on the final posterior matching probability of the two target objects.

[0120] In an optional embodiment, after the posterior matching probability of any two

target objects in multimodality is calculated according to the posterior matching probability function in each modality, the following steps S307 and S308 may be further performed to optimize the posterior matching probability of the any two target objects in multimodality, to improve accuracy of the posterior matching probability of the any two target objects in multimodality.

[0121] Step S307: Determine, for any two target objects, K nearest neighbor objects of the two target objects in each modality, and determine a shared nearest neighbor object of the two target objects.

[0122] In this step, the K nearest neighbor objects of any target object in any modality are obtained by determining, based on similarities between other target objects and the target object in the modality, K target objects having high similarities with the target object by using a KNN algorithm. K may be set and adjusted according to a requirement of an actual application scenario, which is not specifically limited herein. K is a positive integer.

[0123] Further, after K nearest neighbor objects of each target object in each modality are determined, a union set of K nearest neighbor object sets of any target object in respective modalities may be obtained, and the union set is used as a nearest neighbor object set of the target object, including the K nearest neighbor objects of the target object in each modality. For any two target objects, an intersection set of nearest neighbor object sets of the two target objects is determined, and a nearest neighbor object in the intersection set is a shared nearest neighbor object of the two target objects.

[0124] For example, $knn_i^{(m)}$ represents a set of K nearest neighbor objects of a target object i in the mth modality, and a nearest neighbor object set of the target object i is a union set of $knn_i^{(1)}$, $knn_i^{(2)}$, …, and $knn_i^{(M)}$, which may be represented as $knn_i = \{\cup\, knn_i^{(m)}\}$. Similarly, $knn_j^{(m)}$ represents a set of K nearest neighbor objects of a target object j in the mth modality, and a nearest neighbor object set of the target object j is represented as $knn_j = \{\cup\, knn_j^{(m)}\}$. A set of shared nearest neighbor objects of the target object i and the target object j is $knn_{ij} = knn_i \cap knn_j$.

[0125] Step S308: Optimize the posterior matching probability of the two target objects in

multimodality according to posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality.

[0126] In this embodiment, the posterior matching probability of any two target objects in multimodality is optimized according to posterior matching probabilities of the two target objects and a same shared nearest neighbor object in multimodality and base on consistency and a boundary constraint of the posterior matching probability, to improve the accuracy of the posterior matching probability of any two target objects in multimodality.

[0127] A first optimization manner applicable to this step is: calculating a product of the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality, and optimizing the posterior matching probability of the two target objects in multimodality according to a maximum value of the product and the posterior matching probability of the two target objects in multimodality.

[0128] For example, for a posterior matching probability $P(i,j)$ of any two target objects i and j in multimodality obtained through computation, the following formula (12) may be used for optimization:

$$P(i,j) = \max\left(P(i,j), P(i,h) \times P(h,j)\right) \tag{12}$$

[0129] In the formula (12), h refers to any shared nearest neighbor object of the target object i and the target object j, and $h \in knn_{ij}$. $knn_{ij} = knn_i \cap knn_j$ represents a set of shared nearest neighbor objects of the target object i and the target object j. $\max()$ is a function for obtaining the maximum value. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0130] A second optimization manner applicable to this step is: using a sum of the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality as a posterior matching probability sum corresponding to the shared nearest neighbor object; calculating a sum of posterior matching probability sums corresponding to shared nearest neighbor objects to obtain a first summation result; for a first object and a second object of the two target objects, calculating a sum of posterior matching probabilities of the first object and the nearest neighbor objects of the first object in multimodality to obtain a second summation result, calculating a sum of posterior matching probabilities of the second object

and the nearest neighbor objects of the second object in multimodality to obtain a third summation result, and calculating a sum of the second summation result and the third summation result to obtain a fourth summation result; and optimizing the posterior matching probability of the two target objects in multimodality according to a ratio of the first summation result to the fourth summation result.

[0131] For example, for a posterior matching probability $P(i,j)$ of any two target objects i and j in multimodality obtained through computation, the following formula (13) may be used for optimization:

$$P(i,j) = \frac{\sum_{h \in knn_{ij}} (P(i,h)+P(h,j))}{\sum_{h_i \in knn_i} P(i,h_i)+\sum_{h_j \in knn_j} P(j,h_j)} \quad (13)$$

[0132] In the formula (13), h_i refers to any nearest neighbor object in a nearest neighbor object set of the target object i, and $h_i \in knn_i$. h_j refers to any nearest neighbor object in a nearest neighbor object set of the target object j, and $h_j \in knn_j$. Meanings of other characters are the same as those in the foregoing content, and details are not described herein again.

[0133] In this step, the foregoing two optimization manners may further be used to optimize the posterior matching probability of any two target objects in multimodality, and a sequence of performing optimization based on the two optimization manners is not limited. For example, the posterior matching probability of any two target objects in multimodality may be first optimized by using the first optimization manner, and then an optimization result of the first optimization manner is optimized for the second time by using the second optimization manner. Alternatively, the posterior matching probability of any two target objects in multimodality may be first optimized by using the second optimization manner, and then an optimization result of the second optimization manner is optimized for the second time by using the first optimization manner.

[0134] Step S309: Cluster the target objects according to the posterior matching probability of any two target objects in multimodality to generate a clustering result of the target objects.

[0135] After the posterior matching probability of any two target objects in multimodality is obtained, the target objects are clustered based on the posterior matching probability of any two target objects in multimodality, to generate the clustering result of the target objects.

[0136] In this step, any existing clustering algorithm may be used, and the posterior

matching probability of the target objects in multimodality is used as a metric distance for the similarity between the target objects, to perform clustering. A specific clustering procedure is not described herein again.

[0137] In this embodiment, for any two target objects, when feature information of the target objects in one or more modalities is missing, the feature information in the modality(modalities) may be ignored, and the posterior matching probability of the two target objects in remaining multiple modalities is calculated based on feature information of each modality in which no information is missing, so that when feature information in some modalities is missing, clustering can still be accurately performed without supplementing the feature information in the missing modalities. In addition, multimodal feature information is fully used for clustering, importance between different modalities is focused on, and the accuracy and recall rate of the clustering are improved. In addition, T sample pairs in each modality are constructed by using the KNN algorithm, and a posterior matching probability function in each modality is generated through self-learning based on similarities of the T sample pairs in each modality, so that time for generating the posterior matching probability function in each modality through self-learning can be shortened, thereby improving clustering efficiency.

[0138] The multimodal information-based clustering method provided in this application can generate the posterior matching probability function in each single modality through self-learning, thereby enabling unbiased and rapid application to different data distribution scenarios, for example, aggregation and collation of different regional image libraries and clustering of commodities of a same type.

[0139] FIG. 5 is a flowchart of a multimodal information-based clustering method according to another exemplary embodiment of this application. The method provided in this embodiment is applicable to a scenario of clustering and collating different regional image libraries. For images in different regional information libraries captured in different regions, images that belong to a same object are clustered to a same cluster based on geographic location information of image capturing and feature information that is in at least one modality and of an object included in the images. For example, based on images captured in different regions, multimodal clustering is performed on images based on geographic location information of

image capturing, and biometric features of persons included in the images, to aggregate images of a same person to a same cluster. For example, based on motor vehicle images captured in different regions, multimodal clustering is performed on motor vehicle images based on geographic location information of motor vehicle image capturing, license plate numbers, owner identities, and appearance features of motor vehicles included in the motor vehicle images, and the like, to aggregate images of a same motor vehicle into a same cluster.

[0140] As shown in FIG. 5, the method specifically includes the following steps.

[0141] Step S501: Obtain multimodal feature information of to-be-clustered images, where the multimodal feature information includes geographic location information and feature information of an included object in at least one modality.

[0142] The to-be-clustered target objects in this embodiment are images, and multimodal feature information of the images includes geographic location information of image capturing, and feature information of an object included in the images in at least one modality.

[0143] For example, in a scenario of clustering and collating different regional image libraries, the to-be-clustered images may be images of a particular object captured in different regions in a plurality of regional information libraries. Using a human image as an example, multimodal feature information of the image may include location coordinates of image capturing, and features of a human included in the image in at least one of the following modalities: a biological feature, a gait feature, and identity information.

[0144] Step S502: Generate a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the images.

[0145] After the multimodal feature information of the to-be-clustered images is obtained, the posterior matching probability function in each modality is generated through self-learning according to the multimodal feature information of the images. This step may be implemented in a manner similar to step S202 in any one of the foregoing method embodiments. The target object in this step is an image. For details, refer to related content in the foregoing embodiments, and details are not described herein again.

[0146] Step S503: Calculate a posterior matching probability of any two images in multimodality according to the posterior matching probability function in each modality, where the posterior matching probability of two images in multimodality is a probability that the two

images include a same object.

[0147] A specific implementation of this step is similar to the implementation of calculating a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality in the foregoing step S203. The target object in this step is an image. For details, refer to related content in the foregoing embodiments, and details are not described herein again.

[0148] Step S504: Cluster images that include a same object into one category according to the posterior matching probability of any two images in multimodality, to obtain a clustering result of the images.

[0149] After the posterior matching probability of any two images in multimodality is obtained, the images are clustered based on the posterior matching probability of any two images in multimodality, images having a larger posterior matching probability in multimodality are clustered into a same cluster, and images having a smaller posterior matching probability in multimodality are clustered into different clusters, thereby implementing accurate image clustering based on uniqueness of the objects included in the images.

[0150] In this step, any existing clustering algorithm may be used, and the posterior matching probability of the images in multimodality is used as a metric distance for the similarity between the images, to perform clustering. A specific clustering procedure is not described herein again.

[0151] Further, after obtaining the image clustering result, the server sends the image clustering result to the terminal side device, and the terminal side device may execute at least one of the following processing logics according to the image clustering result: information pushing, information searching, or data classification. In another implementation, the server may execute at least one of the following processing logics according to the image clustering result: information pushing, information searching, or data classification, and send an execution result to the terminal side device.

[0152] For example, after obtaining the image clustering result, the server may associate, according to the clustering result, in accordance with clusters to which the images belong, images of a same cluster with an included object, and store the images by classification, to implement the function of information library aggregation and collation.

[0153] For example, after obtaining the image clustering result, the server may further analyze and determine information such as a movement range and a social area of a specified object based on geographic location information of images associated with the specified object.

[0154] In this embodiment, in a scenario of aggregation and collation of different regional information libraries, geographic location information of to-be-clustered images and feature information of an included object in at least one modality are obtained to obtain multimodal feature information of the to-be-clustered images. A posterior matching probability function in each single modality is generated through self-learning, so that the posterior matching probability function in each modality can be automatically learned without using annotation information; a posterior matching probability of any two images in multimodality is calculated according to the posterior matching probability function in each single modality; and the images are clustered according to the posterior matching probability of any two images in multimodality, to generate a clustering result of the images. The posterior matching probability that the images include a same object is calculated based on the multimodal feature information of the images by using the posterior matching probability function generated through self-learning, and is used as a metric distance for the similarity between the images. In comparison with using the Euclidean distance or the cosine similarity between image features in mainstream methods, using the posterior matching probability between the images can measure the similarity between objects included in the images more accurately. Clustering is performed based on the posterior matching probability between the images, which can greatly improve the accuracy and recall rate of the image clustering.

[0155] FIG. 6 is an interaction diagram of a multimodal information-based clustering method according to an exemplary embodiment of this application. As shown in FIG. 6, in the multimodal information-based clustering method, an interaction procedure between a server and a terminal side device is as follows.

[0156] Step S601: The terminal side device obtains multimodal feature information of to-be-clustered target objects.

[0157] In this embodiment, the terminal side device obtains multimodal data of the target objects from a data service device. Data in different modalities may be from different devices. The multimodal data of the target objects is preprocessed to generate multimodal feature

information of the target objects, and the multimodal feature information of the target objects is sent to the server.

[0158] The preprocessing the multimodal data of the target objects may include, but is not limited to, cleaning of dirty data, feature extraction, and the like. Feature information that is required to be obtained varies in different application scenarios, and manners of data preprocessing and feature extraction may be set according to requirements in actual application scenarios. Extraction manners for various types of feature information may be implemented by using common technologies for extracting same types of features, and details are not repeated here.

[0159] Step S602: The terminal side device sends the multimodal feature information of the to-be-clustered target objects to the server.

[0160] Step S603: The server receives the multimodal feature information of the to-be-clustered target objects.

[0161] Step S604: The server generates a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects.

[0162] A specific implementation of this step is the same as the specific implementation of the foregoing step S202. For details, refer to the related content in the foregoing embodiment, which are not herein described again.

[0163] Step S605: The server calculates a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality.

[0164] A specific implementation of this step is the same as the specific implementation of the foregoing step S203. For details, refer to the related content in the foregoing embodiment, which are not herein described again.

[0165] Step S606: The server clusters the target objects according to the posterior matching probability of any two target objects in multimodality to generate a clustering result of the target objects.

[0166] A specific implementation of this step is the same as the specific implementation of the foregoing step S204. For details, refer to the related content in the foregoing embodiment,

which are not herein described again.

[0167] Step S607: The server sends the clustering result of the target objects to the terminal side device.

[0168] Step S608: The terminal side device receives the clustering result of the target objects sent by the server.

[0169] Step S609: The terminal side device executes at least one of the following processing logics according to the clustering result: information pushing, information searching, or data classification.

[0170] For example, in an e-commerce scenario, after an item clustering result is obtained by clustering similar items of different merchants to a same cluster through multimodal clustering, services such as item searching and item recommendation may be implemented based on multimodal information such as videos, pictures, introduction texts, and price information.

[0171] For example, according to an item currently browsed by a user, information about similar items are searched for, and the information about the similar items is presented/pushed to the user to facilitate the user's selection of desired items. Alternatively, according to an image submitted by a user, an item included in the image and information about similar items thereof are searched for, and the item included in the image and the information about similar items thereof are presented/pushed to the user to facilitate the user's selection of desired items.

[0172] For example, in a scenario of clustering and collating different regional information libraries, among the images captured in different regions, images that belong to a same object are clustered to a same cluster based on geographic location information of image capturing and feature information of objects included in the images in at least one modality. Further, according to a clustering result, images of a same cluster are associated with the included object according to the cluster to which the images belong, and the images are stored by classification to implement the function of information library aggregation and collation.

[0173] This embodiment provides the interaction procedure between the server and the terminal side device in the multimodal information-based clustering method.

[0174] FIG. 7 is a schematic structural diagram of a multimodal information-based clustering apparatus according to an exemplary embodiment of this application. The

multimodal information-based clustering apparatus provided in this embodiment of this application can execute the processing procedure provided in the multimodal information-based clustering method embodiment. As shown in FIG. 7, the multimodal information-based clustering apparatus 70 includes: a multimodal feature obtaining module 71, a self-learning module 72, a posterior matching probability calculating module 73, and a multimodal clustering module 74.

[0175] Specifically, the multimodal feature obtaining module 71 is configured to obtain multimodal feature information of to-be-clustered target objects.

[0176] The self-learning module 72 generates a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects.

[0177] The posterior matching probability calculating module 73 calculates a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality.

[0178] The multimodal clustering module 74 clusters the target objects according to the posterior matching probability of any two target objects in multimodality, to generate a clustering result of the target objects.

[0179] In an optional embodiment, when implementing generating the posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects, the self-learning module 72 is further configured to:

calculate a similarity between any two target objects in a same modality according to feature information of the target objects in the same modality; construct a plurality of sample pairs in each modality according to a similarity between any two target objects in each modality, where the sample pair includes two target objects; and generate the posterior matching probability function in each modality through self-learning by using an optimization algorithm and according to similarities of the plurality of sample pairs in each modality.

[0180] In an optional embodiment, when implementing constructing the plurality of sample pairs in each modality according to the similarity between any two target objects in each modality, where the sample pair includes two target objects, the self-learning module 72 is further configured to:

determine, for any modality, K nearest neighbor objects of each of the target objects in the modality according to a similarity between any two target objects in the modality, where K is a positive integer; and form a sample pair by using each of the target objects and the K nearest neighbor objects in the modality respectively, to obtain a plurality of sample pairs in the modality.

[0181] In an optional embodiment, when implementing generating the posterior matching probability function in each modality through self-learning by using the optimization algorithm and according to the similarities of the plurality of sample pairs in each modality, the self-learning module 72 is further configured to:

initialize the posterior matching probability function in each modality to a monotonically increasing piecewise function, where the piecewise function includes to-be-solved piecewise parameters; construct an objective function based on the posterior matching probability function in each modality, where the objective function includes the to-be-solved piecewise parameters; minimize the objective function by using the optimization algorithm and according to the similarities of the plurality of sample pairs in each modality to solve and obtain the segment parameters; and substitute the segment parameters into the piecewise function to obtain the posterior matching probability function in each modality.

[0182] In an optional embodiment, when implementing constructing the objective function based on the posterior matching probability function in each modality, the self-learning module 72 is further configured to:

construct a posterior matching probability function in multimodality according to the posterior matching probability function in each modality; construct a consistency loss function and a constraint loss function according to the posterior matching probability function in each modality and the posterior matching probability function in multimodality; and

construct the objective function according to the consistency loss function and the constraint loss function.

[0183] In an optional embodiment, when implementing calculating the posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality, the posterior matching probability calculating module 73 is further configured to:

determine, according to a similarity between any two target objects in any modality, a posterior matching probability of the any two target objects in the modality by using a posterior matching probability function in the modality; and calculate the posterior matching probability of the any two target objects in multimodality according to the posterior matching probability of the any two target objects in each modality.

[0184] In an optional embodiment, after calculating the posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality, the posterior matching probability calculating module 73 is further configured to:

determine, for any two target objects, K nearest neighbor objects of the two target objects in each modality, and determine a shared nearest neighbor object of the two target objects, where K is a positive integer; and optimize a posterior matching probability of the two target objects in multimodality according to posterior matching probabilities of the two target objects and a same shared nearest neighbor object in multimodality.

[0185] In an optional embodiment, when implementing optimizing the posterior matching probability of the two target objects in multimodality according to the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality, the posterior matching probability calculating module 73 is further configured to:

calculate a product of the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality, and optimize the posterior matching probability of the two target objects in multimodality according to a maximum value of the product and the posterior matching probability of the two target objects in multimodality.

[0186] In an optional embodiment, when implementing optimizing the posterior matching probability of the two target objects in multimodality according to the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality, the posterior matching probability calculating module 73 is further configured to:

use a sum of posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality as a posterior matching probability sum

corresponding to the shared nearest neighbor object; calculate a sum of posterior matching probability sums corresponding to shared nearest neighbor objects to obtain a first summation result; for a first object and a second object of the two target objects, calculate a sum of posterior matching probabilities of the first object and the nearest neighbor objects of the first object in multimodality to obtain a second summation result; calculate a sum of posterior matching probabilities of the second object and the nearest neighbor objects of the second object in multimodality to obtain a third summation result; and calculate a sum of the second summation result and the third summation result to obtain a fourth summation result; and optimize the posterior matching probability of the two target objects in multimodality according to a ratio of the first summation result to the fourth summation result.

[0187] In an optional embodiment, when implementing obtaining of the multimodal feature information of to-be-clustered target objects, the multimodal feature obtaining module 71 is further configured to receive the multimodal feature information of the to-be-clustered target objects sent by a terminal side device.

[0188] After generating the clustering result of the target objects, the multimodal clustering module 74 is further configured to output the clustering result of the target objects to the terminal side device.

[0189] The apparatus provided in this embodiment of this application may be specifically configured to perform the processing procedure performed by the server in any one of the foregoing method embodiments. A specific function and a technical effect that can be achieved are not described herein again.

[0190] FIG. 8 is a schematic structural diagram of a server according to an embodiment of this application. As shown in FIG. 8, the server includes a memory 801 and a processor 802. The memory 801 is configured to store computer-executable instructions, and may be configured to store various other data to support operations on the server. The processor 802 is communicatively connected to the memory 801 and is configured to execute the computer-executable instructions stored in the memory 801 to implement the technical solutions provided in any one of the foregoing method embodiments. Specific functions and technical effects that can be implemented are similar, and details are not described herein again. In FIG. 8, an example in which the server is a cloud server is used for description, and the server may

alternatively be a locally deployed server.

[0191] Optionally, as shown in FIG. 8, the server further includes other components such as a firewall 803, a load balancer 804, a communication component 805, and a power supply component 806. Only some components are schematically provided in FIG. 8, and it does not mean that the server includes only the components shown in FIG. 8.

[0192] An embodiment of this application further provides a computer-readable storage medium. The computer-readable storage medium has computer-executable instructions stored therein, and the computer-executable instructions, when executed by a processor, are used to implement the solution provided in any one of the foregoing method embodiments. Specific functions and technical effects that can be implemented are not described here again.

[0193] An embodiment of this application further provides a computer program product. The computer program product includes a computer program. The computer program is stored in a readable storage medium. At least one processor of a server may read the computer program from the readable storage medium, and at least one processor executes the computer program, so that the server executes the solution provided in any one of the foregoing method embodiments. Specific functions and technical effects that can be implemented are not described here again. An embodiment of this application provides a chip, including: a processing module and a communication interface. The processing module can execute the technical solution of the server in the foregoing method embodiments. Optionally, the chip further includes a storage module (for example, a memory). The storage module is configured to store instructions, the processing module is configured to execute the instructions stored in the storage module, and execution of the instructions stored in the storage module causes the processing module to perform the technical solution provided in any one of the foregoing method embodiments.

[0194] The memory may be an object storage service (OSS).

[0195] The memory may be implemented by any type of volatile or non-volatile storage device or a combination thereof, such as a static random access memory (SRAM), an electrically erasable programmable read-only memory (EEPROM), an erasable programmable read-only memory (EPROM), a programmable read-only memory (PROM), a read-only memory (ROM), a magnetic memory, a flash memory, a magnetic disk, or an optical disc.

[0196] The communication component is configured to facilitate wired or wireless communication between a device on which the communication component is located and another device. The device on which the communication component is located may access a wireless network based on a communication standard, such as a mobile communication network such as a mobile hotspot (Wi-Fi), a second generation mobile communication system (2G), a third generation mobile communication system (3G), a fourth generation mobile communication system (4G)/long term evolution (LTE), or a fifth generation mobile communication system (5G), or a combination thereof. In an exemplary embodiment, the communication component receives a broadcast signal or broadcast related information from an external broadcast management system through a broadcast channel. In an exemplary embodiment, the communication component further includes a near field communication (NFC) module to facilitate short range communication. For example, the NFC module may be implemented based on a radio frequency identification (RFID) technology, an infrared data association (IrDA) technology, an ultra-wideband (UWB) technology, a Bluetooth (BT) technology, and other technologies.

[0197] The power supply component supplies power to various components of the device on which the power supply component is located. The power supply component may include a power supply management system, one or more power supplies, and other components associated with generating, managing, and allocating power for the device on which the power supply component is located.

[0198] A person skilled in the art should understand that the embodiments of the present invention may be provided as a method, a system, or a computer program product. Therefore, the present invention may use a form of hardware-only embodiments, software-only embodiments, or embodiments with a combination of software and hardware. Moreover, the present invention may use a form of a computer program product that is implemented on one or more computer-usable storage media (including but not limited to a disk memory, a compact disc read-only memory (CD-ROM), an optical memory, and the like) that include computer-usable program code.

[0199] The present invention is described with reference to the flowcharts and/or block diagrams of the method, the device (system), and the computer program product according to

the embodiments of the present invention. It should be understood that computer program instructions may be used to implement each process and/or each block in the flowcharts and/or the block diagrams and a combination of a process and/or a block in the flowcharts and/or the block diagrams. These computer program instructions may be provided for a general-purpose computer, a special-purpose computer, an embedded processor, or a processor of any other programmable data processing device to generate a machine, so that the instructions executed by the computer or the processor of any other programmable data processing device generate an apparatus for implementing a specific function in one or more procedures in the flowcharts and/or in one or more blocks in the block diagrams.

[0200] These computer program instructions may alternatively be stored in a computer-readable memory that can instruct the computer or any other programmable data processing device to work in a specific manner, so that the instructions stored in the computer-readable memory generate an artifact that includes an instruction apparatus. The instruction apparatus implements a specific function in one or more processes in the flowcharts and/or in one or more blocks in the block diagrams.

[0201] These computer program instructions may also be loaded onto a computer or another programmable data processing device, so that a series of operational steps are performed on the computer or the another programmable device, thereby generating computer-implemented processing. Therefore, the instructions executed on the computer or the another programmable device provide operations for implementing a specific function in one or more procedures in the flowcharts and/or in one or more blocks in the block diagrams.

[0202] In a typical configuration, a computing device includes one or more processors (CPUs), an input/output interface, a network interface, and a memory.

[0203] The memory may include a form such as a volatile memory, a random-access memory (RAM) and/or a non-volatile memory such as a read-only memory (ROM) or a flash RAM in computer-readable media. The memory is an example of a computer-readable medium.

[0204] The computer-readable medium includes a non-volatile medium and a volatile medium, a removable medium and a non-removable medium, which may implement storage of information by using any method or technology. The information may be a computer-readable instruction, a data structure, a program module, or other data. Examples of a storage medium

of a computer includes, but is not limited to, a phase-change memory (PRAM), a static random access memory (SRAM), a dynamic random access memory (DRAM), or other types of random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EEPROM), a flash memory or another storage technology, a compact disc read-only memory (CD-ROM), a digital versatile disc (DVD) or another optical storage, a cartridge tape, a magnetic tape, a magnetic disk storage or another magnetic storage device, or any other non-transmission medium, which may be configured to store information accessible by a computing device. According to limitations of this specification, the computer-readable medium does not include transitory computer-readable media (transitory media), such as a modulated data signal and a modulated carrier.

[0205] It should be noted that, user information (including but not limited to user equipment information, user attribute information, and the like) and data (including but not limited to data for analysis, stored data, displayed data, and the like) in this application are all information and data that are authorized by a user or that are fully authorized by all parties, and related data needs to be collected, used, and processed by complying with relevant laws, regulations, and standards. In addition, a corresponding operation entry is provided for the user to select to authorize or reject.

[0206] In addition, some procedures described in the foregoing embodiments and accompanying drawings include a plurality of operations performed in a particular order. However, it should be clearly learned that these operations may not be performed in the order in which these operations are performed in the specification or may be performed concurrently. The sequence numbers are merely used to distinguish between different operations, and do not represent any execution order. In addition, these procedures may include more or fewer operations, and these operations may be performed in an order or may be performed concurrently. It should be noted that, descriptions of “first”, “second”, and the like in the specification are used to distinguish between different messages, devices, modules, and the like, and do not indicate a sequence and do not limit that “first” and “second” are different types either. “A plurality of” means two or more, unless otherwise explicitly and specifically defined.

[0207] It should be noted that, user information (including but not limited to user equipment information, user personal information, and the like) and data (including but not limited to data

for analysis, stored data, displayed data, and the like) in this application are all information and data that are authorized by a user or that are fully authorized by all parties, and related data needs to be collected, used, and processed by complying with relevant laws, regulations, and standards. In addition, a corresponding operation entry is provided for the user to select to authorize or reject.

[0208] After considering the specification and practicing the present disclosure, a person skilled in the art may easily conceive of other implementations of this application. This application is intended to cover any variations, uses, or adaptive changes of this application. These variations, uses, or adaptive changes follow the general principles of this application and include common general knowledge or common technical means in the art, which are not disclosed in this application. The specification and the embodiments are considered as merely illustrative, and the scope and spirit of this application are indicated in the following claims.

[0209] It should be understood that this application is not limited to the precise structures described above and shown in the accompanying drawings, and various modifications and changes can be made without departing from the scope of this application. The scope of this application is subject only to the appended claims.

CLAIMS

1. A multimodal information-based clustering method, comprising:

obtaining multimodal feature information of to-be-clustered target objects;

generating a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects;

calculating a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality; and

clustering the target objects according to the posterior matching probability of any two target objects in multimodality, to generate a clustering result of the target objects.

2. The method according to claim 1, wherein the generating the posterior matching probability function in each modality through self-learning according to the multimodal feature information of the target objects comprises:

calculating, according to feature information of the target objects in a same modality, a similarity between any two target objects in the same modality;

constructing a plurality of sample pairs in each modality according to a similarity between any two target objects in each modality, wherein each of the plurality of sample pairs comprises two target objects; and

generating the posterior matching probability function in each modality through self-learning by using an optimization algorithm and according to similarities of the plurality of sample pairs in each modality.

3. The method according to claim 2, wherein the constructing the plurality of sample pairs in each modality according to the similarity between any two target objects in each modality comprises:

determining, for any modality, K nearest neighbor objects of each of the target objects in the modality according to a similarity between any two target objects in the modality, wherein K is a positive integer; and

forming a sample pair by using each of the target objects and the K nearest neighbor objects in the modality respectively, to obtain a plurality of sample pairs in the modality.

4. The method according to claim 2, wherein the generating the posterior matching probability function in each modality through self-learning by using the optimization algorithm and according to the similarities of the plurality of sample pairs in each modality comprises:

initializing the posterior matching probability function in each modality to a monotonically increasing piecewise function, wherein the piecewise function comprises to-be-solved segment parameters;

constructing an objective function based on the posterior matching probability function in each modality, wherein the objective function comprises the to-be-solved segment parameters;

minimizing the objective function by using the optimization algorithm and according to the similarities of the plurality of sample pairs in each modality, to solve and obtain the segment parameters; and

substituting the segment parameters into the piecewise function to obtain the posterior matching probability function in each modality.

5. The method according to claim 4, wherein the constructing the objective function based on the posterior matching probability function in each modality comprises:

constructing a posterior matching probability function in multimodality according to the posterior matching probability function in each modality;

constructing a consistency loss function and a constraint loss function according to the posterior matching probability function in each modality and the posterior matching probability function in multimodality; and

constructing the objective function according to the consistency loss function and the constraint loss function.

6. The method according to claim 1, wherein the calculating the posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality comprises:

determining, according to a similarity between any two target objects in any modality, a posterior matching probability of the any two target objects in the modality by using a posterior matching probability function in the modality; and

calculating the posterior matching probability of the any two target objects in multimodality according to the posterior matching probability of the any two target objects in each modality.

7. The method according to any one of claims 1 to 6, wherein after the calculating a posterior matching probability of any two target objects in multimodality according to the posterior matching probability function in each modality, the method further comprises:

determining, for any two target objects, K nearest neighbor objects of the two target objects in each modality, and determining a shared nearest neighbor object of the two target objects, wherein K is a positive integer; and

optimizing the posterior matching probability of the two target objects in multimodality according to posterior matching probabilities of the two target objects and a same shared nearest neighbor object in multimodality.

8. The method according to claim 7, wherein the optimizing the posterior matching probability of the two target objects in multimodality according to posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality comprises:

calculating a product of the posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality, and optimizing the posterior matching probability of the two target objects in multimodality according to a maximum value of the product and the posterior matching probability of the two target objects in multimodality.

9. The method according to claim 7, wherein the optimizing the posterior matching probability of the two target objects in multimodality according to posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality comprises:

using a sum of posterior matching probabilities of the two target objects and the same shared nearest neighbor object in multimodality as a posterior matching probability sum corresponding to the shared nearest neighbor object;

calculating a sum of posterior matching probability sums corresponding to shared nearest neighbor objects, to obtain a first summation result;

for a first object and a second object in the two target objects, calculating a sum of posterior matching probabilities of the first object and the nearest neighbor objects of the first object in multimodality, to obtain a second summation result; calculating a sum of posterior matching probabilities of the second object and the nearest neighbor objects of the second object in multimodality, to obtain a third summation result; and calculating a sum of the second summation result and the third summation result, to obtain a fourth summation result; and

optimizing the posterior matching probability of the two target objects in multimodality according to a ratio of the first summation result to the fourth summation result.

10. The method according to any one of claims 1 to 6, wherein the obtaining the multimodal feature information of the to-be-clustered target objects comprises:

receiving the multimodal feature information of the to-be-clustered target objects sent by a terminal side device; and

after generating the clustering result of the target objects, the method further comprises:

outputting the clustering result of the target objects to the terminal side device.

11. A multimodal information-based clustering method, comprising:

obtaining multimodal feature information of to-be-clustered images, wherein the multimodal feature information comprises geographic location information and feature information of a comprised object in at least one modality;

generating a posterior matching probability function in each modality through self-learning according to the multimodal feature information of the images;

calculating a posterior matching probability of any two images in multimodality according to the posterior matching probability function in each modality, wherein the posterior matching probability of two images in multimodality is a probability that the two images comprise a same object; and

clustering images that comprise a same object into one category according to the posterior matching probability of any two images in multimodality, to obtain a clustering result of the images.

12. A multimodal information-based clustering method, applied to a terminal side device, and comprising:

obtaining multimodal feature information of to-be-clustered target objects;

sending the multimodal feature information of the to-be-clustered target objects to a server;

receiving a clustering result of the target objects sent by the server, wherein the clustering result of the target objects is generated by using the method according to any one of claims 1 to 10; and

executing at least one of the following processing logics according to the clustering result: information pushing, information searching, or data classification.

13. A server, comprising:

a processor and a memory communicatively connected to the processor, wherein

the memory stores computer-executable instructions; and

the processor executes the computer-executable instructions stored in the memory to implement the method according to any one of claims 1 to 11.

14. A computer-readable storage medium, wherein the computer-readable storage medium has computer-executable instructions stored therein, and the computer-executable instructions, when executed by a processor, are used to implement the method according to any of claims 1 to 11.