CLAIMS

1. A method for processing model parameters, comprising:

obtaining a to-be-sharded model parameter set, wherein the model parameter set comprises a multi-dimensional array corresponding to a feature embedding;

obtaining attribute information of a storage system storing the to-be-sharded model parameter set, wherein the storage system storing the to-be-sharded model parameter set is different from a system on which a model corresponding to the to-be-sharded model parameter set operates; and

storing the to-be-sharded model parameter set in the storage system based on the attribute information.

2. The method according to claim 1, wherein the obtaining a to-be-sharded model parameter set comprises:

obtaining a model file corresponding to the to-be-sharded model parameter set;

performing analysis on the model file to obtain an analysis graph corresponding to the model file, wherein the analysis graph comprises variables and operations on the variables;

extracting a variable from the analysis graph as a target variable based on a preset field; and

obtaining the to-be-sharded model parameter set based on a feature embedding corresponding to the target variable.

3. The method according to claim 2, wherein the obtaining the to-be-sharded model parameter set based on a feature embedding corresponding to the target variable comprises:

obtaining a data amount of a multi-dimensional array corresponding to the feature embedding corresponding to the target variable; and

in response to determining that the obtained data amount is greater than a preset data amount threshold, obtaining the multi-dimensional array corresponding to the feature

embedding corresponding to the target variable as the to-be-sharded model parameter set.

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

receiving a feature embedding query operation corresponding to the to-be-sharded model parameter set;

based on the feature embedding query operation, querying the storage system for a result matching the feature embedding query operation; and

transmitting the result matching the feature embedding query operation.

5. The method according to any one of claims 1 to 4, wherein

the attribute information comprises storage space distribution of the storage system; and

the storing the to-be-sharded model parameter set in the storage system based on the attribute information comprises:

generating a model parameter allocation table based on the attribute information, wherein the model parameter allocation table represents a correspondence relationship between the to-be-sharded model parameter set and storage locations of the storage system; and

based on the model parameter allocation table, storing the to-be-sharded model parameter set in a corresponding storage location in the storage system, and generating a query interface.

6. The method according to any one of claims 1 to 3, wherein

the attribute information comprises an identifier of a key-value storage system; and

the storing the to-be-sharded model parameter set in the storage system based on the attribute information comprises:

obtaining a previously updated model parameter set as a reference model parameter set;

determining a difference between the to-be-sharded model parameter set and the reference model parameter set, and generating an updated model parameter set;

storing the updated model parameter set in the storage system based on the attribute information.

7. The method according to claim 6, further comprising:

receiving a feature embedding query operation corresponding to the to-be-sharded model parameter set;

generating a corresponding number of remote procedure call requests based on the feature embedding query operation;

transmitting the remote procedure call requests to the key-value storage system; and

in response to receiving response information from the key-value storage system comprising a result that matches the feature embedding query operation, transmitting the result that matches the feature embedding query operation.

8. An apparatus for processing model parameters, comprising:

a first obtaining unit, configured to obtain a to-be-sharded model parameter set, wherein the model parameter set comprises a multi-dimensional array corresponding to a feature embedding;

a second obtaining unit, configured to obtain attribute information of a storage system storing the to-be-sharded model parameter set, wherein the storage system storing the to-be-sharded model parameter set is different from a system on which a model corresponding to the to-be-sharded model parameter set operates; and

a storage unit, configured to store the to-be-sharded model parameter set in the storage system based on the attribute information.

9. A server, comprising:

one or more processors; and

a storage device, storing one or more programs; wherein

the one or more programs, when executed by the one or more processors, cause the one or more processors to perform the method according to any one of claims 1 to 7.

10. A computer-readable medium storing a computer program, wherein the computer program, when executed by a processor, causes the processor to perform the method according to any one of claims 1 to 7.