

VECTOR RETRIEVAL METHODS AND APPARATUSES, DEVICES, AND STORAGE MEDIA

ABSTRACT

One or more embodiments of this application provide a vector retrieval method and apparatus,
5 a device, and a storage medium. A graph index constructed based on an original vector and a
compression vector corresponding to the original vector are stored in a first-level storage resource;
a node in the graph index is associated with a compression vector, and an edge represents that a
near neighbor relationship exists between nodes connected to the edge; the original vector is stored
in a second-level storage resource; data access performance of the first-level storage resource is
10 better than that of the second-level storage resource; and the method includes: in response to a
query vector, obtaining, from the first-level storage resource, at least one compression vector
having a highest similarity with the query vector based on the graph index, and after obtaining is
completed, obtaining at least one original vector corresponding to the at least one compression
vector from the second-level storage resource; and separately calculating a similarity between the
15 query vector and the at least one original vector, sorting the at least one original vector based on
the calculated similarity, and determining, based on the sorting result, a vector retrieval result
corresponding to the query vector.

FIG. 5

20