(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114722244 A
(43)申请公布日 2022.07.08

(21)申请号 202210473409.X

(22)申请日 2022.04.29

(71)申请人 上海徐毓智能科技有限公司
地址 200082 上海市杨浦区荆州路168号安联大厦1601

(72)发明人 谢超 程倩雅 许维芷 易小萌

(74)专利代理机构 上海华诚知识产权代理有限公司 31300
专利代理师 肖华

(51)Int.Cl.
G06F 16/901(2019.01)
G06F 16/9032(2019.01)
G06F 16/906(2019.01)

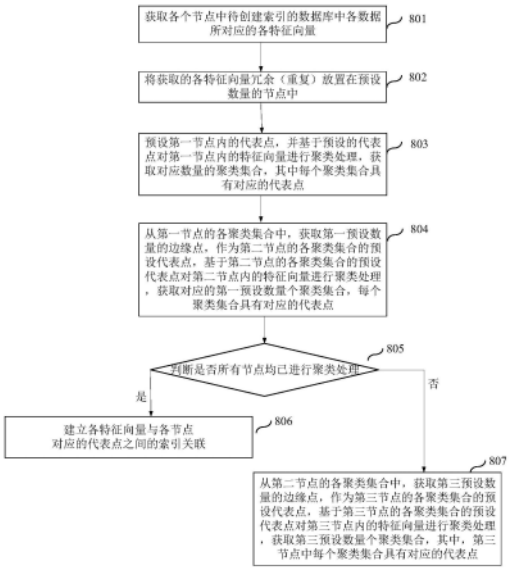
权利要求书3页 说明书31页 附图16页

(54)发明名称

索引构建方法、装置、数据系统及搜索方法

(57)摘要

本申请涉及数据检索技术领域，公开了一种索引构建方法、装置、数据系统及搜索方法。其中，索引构建方法包括：确定当前节点的各聚类集合；在当前节点的各聚类集合中确定第一预设数量的边缘点作为下一节点的预设代表点；基于下一节点的预设代表点，获取下一节点的各聚类集合；确定目标向量以及目标向量在各节点内所在的第一聚类集合；分别在各节点中，建立目标向量与目标向量所在的第一聚类集合之间的索引关联。基于上述方案，可以使得向量检索的过程中，搜索各节点中目标向量所在的聚类集合，能够有效避免由于聚类集合的代表点并不能完全表示聚类集合中的所有特征向量所引起的搜索精确度较低的问题，进而有效提高搜索精度。



CN 114722244 A

1.一种索引构建方法,应用于电子系统,其特征在于,所述电子系统包括用于向量存储的多个节点;

所述方法包括:

确定当前节点的各聚类集合;

在所述当前节点的各聚类集合中确定第一预设数量的边缘点作为下一节点的预设代表点;

基于所述下一节点的预设代表点,获取所述下一节点的各聚类集合;

确定目标向量以及所述目标向量在各节点内所在的第一聚类集合;

分别在所述各节点中,建立所述目标向量与所述目标向量所在的第一聚类集合之间的索引关联。

2.根据权利要求1所述的索引构建方法,其特征在于,其中,确定所述目标向量在各节点内所在的第一聚类集合,包括:

获取各聚类集合对应的预设代表点;

获取所述目标向量与各节点内各聚类集合对应的预设代表点之间的距离;

将所述各节点内各聚类集合中,对应的预设代表点与所述目标向量之间的距离最近的聚类集合作为所述目标向量所在的第一聚类集合。

3.根据权利要求2所述的索引构建方法,其特征在于,其中,在所述当前节点的各聚类集合中确定第一预设数量的边缘点,包括:

根据所述第一预设数量以及所述当前节点内各聚类集合的特征向量的数量,确定所述当前节点内各聚类集合的边缘点的第二预设数量;

从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点。

4.根据权利要求3所述的索引构建方法,其特征在于,其中,从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点,包括:

确定所述当前节点内,每个聚类集合的相邻聚类集合;

对于对应聚类集合的各特征向量,获取所述各特征向量与所述对应聚类集合的代表点之间的第一距离,并获取所述各特征向量与所述相邻聚类集合的代表点之间的第二距离;

根据所述第一距离和所述第二距离,确定对应的第一序列,将所述第一序列中的第二预设数量的特征向量作为所述对应聚类集合的边缘点。

5.根据权利要求4所述的索引构建方法,其特征在于,其中,根据所述第一距离和所述第二距离,确定第一序列,将所述第一序列中的第二预设数量的特征向量作为所述对应聚类集合的边缘点,包括:

获取所述第一距离和所述第二距离的距离之和;

对所述各特征向量,根据所述距离之和由大至小排序,获取所述第一序列;

将所述第一序列中的前第二预设数量的特征向量作为所述对应聚类集合的边缘点。

6.根据权利要求4所述的索引构建方法,其特征在于,其中,根据所述第一距离和所述第二距离,确定第一序列,将所述第一序列中的第二预设数量的特征向量作为所述对应聚类集合的边缘点,包括:

获取所述第二距离和所述第一距离的差值与所述第一距离的比值;

对所述各特征向量,根据所述比值由小至大排序,获取所述第一序列;

将所述第一序列中的前第二预设数量的特征向量作为所述对应聚类集合的边缘点。

7.根据权利要求3所述的索引构建方法，其特征在于，其中，从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点，包括：

确定对应聚类集合中各特征向量与所述对应集合的代表点之间的第一距离；

根据所述第一距离，从所述各特征向量中，确定第一序列，将所述第一序列中的第二预设数量的特征向量作为所述第一聚类集合的边缘点。

8.根据权利要求3所述的索引构建方法，其特征在于，其中，从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点，包括：

确定所述当前节点内，每个聚类集合的相邻聚类集合；

确定对应聚类集合中各特征向量与所述对应集合的相邻聚类集合的代表点之间的第二距离；

根据所述第二距离，从所述各特征向量中，确定第一序列，将所述第一序列中的第二预设数量的特征向量作为所述第一聚类集合的边缘点。

9.根据权利要求1-8任一项所述的方法，其特征在于，所述距离包括欧式距离、内积距离和汉明距离。

10.一种索引构建装置，其特征在于，包括：

第一确定单元，用于确定目标向量以及所述目标向量在当前节点内所在的第一聚类集合，所述第一聚类集合具有第一代表点；

第二确定单元，用于确定所述目标向量在下一节点内所在的第二聚类集合，所述第二聚类集合具有对应的代表点，所述第二聚类集合对应的预设代表点为所述第一节点内其中一个聚类集合的边缘点；其中，所述第一节点内的特征向量与所述第二节点内的特征向量相同；

关联单元，分别在各所述节点中，建立所述目标向量与所述目标向量所在的第一聚类集合之间的索引关联。

11.一种向量搜索方法，其特征在于，包括：

获取查询向量；

获取各节点中各聚类集合的代表点与查询向量的第三距离；

根据所述第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离；

根据所述第四距离，确定出所述查询向量在所述各节点中对应的第一目标向量；

根据所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量。

12.根据权利要求11所述的向量搜索方法，其特征在于，所述按照所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量，包括：

根据所述第四距离，对所述查询向量在所述各节点中对应的第一目标向量，按照所述第四距离由小至大排序，确定前设定次序的目标向量为所述查询向量对应的第二目标向量。

13.一种搜索装置，其特征在于，应用于数据系统：包括：

第一获取单元，用于获取查询向量；

第二获取单元，用于获取各节点中各聚类集合的代表点与查询向量的第三距离；

第一确定单元，用于根据所述第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第二距离；

第二确定单元，用于根据所述第四距离，确定出所述查询向量在所述各节点中对应的第一目标向量；

第三确定单元，用于根据所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量。

14.一种检索系统，其特征在于，包括权利要求10所述的索引构建装置，和/或权利要求13所述的搜索装置。

15.一种索引结构，其特征在于，包括多个子索引结构，每个子索引结构包括多个代表点项和多个倒排文件项；所述代表点项包括数据系统中各个聚类集合所对应的各代表点，所述倒排文件项包括所述各代表点所对应的各倒排文件；

所述每个倒排文件包括所述每个倒排文件对应的代表点所对应的聚类集合中的各个特征向量。

16.一种电子设备，其特征在于，包括：存储器，用于存储由电子设备的一个或多个处理器执行的指令，以及处理器，是所述电子设备的所述一个或多个处理器之一，用于执行权利要求1至9中任一项所述的索引构建方法或者权利要求11至12中任一项所述的搜索方法。

17.一种可读介质，其特征在于，所述可读介质上存储有指令，该指令在电子设备上执行时使机器执行权利要求1至9中任一项所述的索引构建方法或者权利要求11至12中任一项所述的搜索方法。

18.一种计算机程序产品，其特征在于，包括指令，所述指令用于实现权利要求1至9中任一项所述的索引构建方法或者权利要求11至12中任一项所述的搜索方法。

索引构建方法、装置、数据系统及搜索方法

技术领域

[0001] 本申请涉及数据检索技术领域，特别涉及一种索引构建方法、装置、数据系统及搜索方法。

背景技术

[0002] 随着数据的快速增长，数据检索广泛应用于图像、视频、语音、蛋白质分子结构检索等领域中，由于各种数据，例如图片数据等均可以被抽象为高维度的特征向量，因此数据之间的相似度可以被量化为向量空间中的特征向量之间的距离。例如，两个特征向量之间的距离越近，则该两个特征向量对应的原始数据的相似度越高。因此数据检索可以转化为在向量空间中的向量搜索，即将在数据库中搜索与待查询数据相似的若干个数据的过程，转化为在数据库中搜索距离待查询数据所对应的查询向量最近的若干个特征向量的过程。

[0003] 目前，一些检索系统会为数据库构建倒排索引以便于用户检索。其中，倒排索引的构建方法为首先将数据库中各数据对应的各特征向量通过聚类处理，例如通过k-means聚类处理，把整个向量空间划分为若干个聚类集合，每个聚类集合具有对应的代表点，并将每个特征向量归入到距离自身最近的代表点所对应的聚类集合中。如此，在进行查询向量的检索时，系统将会根据查询向量与多个代表点的距离确定出与查询向量最近的代表点，并对该代表点所在的聚类集合中的所有的特征向量进行搜索，可以理解，对所有的特征向量进行搜索即是获取每个特征向量与查询向量之间的距离。然后将该聚类集合中与查询向量之间距离较近的若干个特征向量作为搜索结果。

[0004] 但是，由于检索系统是确定与查询向量距离最近的代表点，然后仅仅将此代表点所在的聚类集合内与查询向量距离较近的若干个特征向量作为搜索结果，所以会存在，在其他代表点所在的聚类集合内，虽然代表点距离查询向量远，反而存在与查询向量距离更近的若干个特征向量的情况，使得未获取到更加准确的查询向量的目标向量，导致检索结果的精确度较低。

发明内容

[0005] 为解决向量检索方法检索结果的精确度较低的问题，本申请实施例提供了一种索引构建方法、装置、数据系统及搜索方法。

[0006] 第一方面，本申请实施例提供了一种索引构建方法，应用于电子系统，其中，所述电子系统包括用于向量存储的多个节点，所述方法包括：

[0007] 确定当前节点的各聚类集合；

[0008] 在所述当前节点的各聚类集合中确定第一预设数量的边缘点作为下一节点的预设代表点；

[0009] 基于所述下一节点的预设代表点，获取所述下一节点的各聚类集合；

[0010] 确定目标向量以及所述目标向量在各节点内所在的第一聚类集合；

[0011] 分别在所述各节点中，建立所述目标向量与所述目标向量所在的第一聚类集合之

间的索引关联。

[0012] 可以理解的，本申请实施例提供的索引构建方法，将当前节点的各聚类集合中确定第一预设数量的边缘点作为下一节点的预设代表点，基于下一节点的预设代表点，获取下一节点的各聚类集合。如此，在检索的过程中，通过获取查询向量距离各节点内的聚类集合的代表点即中心点的距离，并可以根据获取的查询向量与各节点的代表点之间的距离，获取查询向量在各节点内对应的目标向量。然后可以将查询向量在各节点内对应的目标向量汇总进行对比，以实现搜索到与查询向量距离最近的，或最为精确的目标向量，能够有效提高搜索精度。

[0013] 可以理解，本申请实施例中所提及的第一聚类集合中的目标向量可以为第一聚类集合中的任一向量。

[0014] 可以理解，上述边缘点可以为第一聚类集合中的边缘特征向量，即距离第一聚类集合对应的代表点较远的特征向量。

[0015] 可以理解，本申请实施例中，当前节点可以为任一节点，例如第一节点、第二节点，以及第二节点的后续节点。假设当前节点为第一节点，则下一节点即为第二节点。

[0016] 在上述第一方面的一种可能的实现中，其中，确定所述目标向量在各节点内所在的第一聚类集合，包括：

[0017] 获取各聚类集合对应的预设代表点；

[0018] 获取所述目标向量与各节点内各聚类集合对应的预设代表点之间的距离；

[0019] 将所述各节点内各聚类集合中，对应的预设代表点与所述目标向量之间的距离最近的聚类集合作为所述目标向量所在的第一聚类集合。

[0020] 可以理解，本申请实施例中根据目标向量与各节点内各聚类集合对应的预设代表点之间的距离，并在各节点内各聚类集合中，对应的预设代表点与目标向量之间的距离最近的聚类集合作为目标向量所在的第一聚类集合，便于更加方便的根据代表点与查询向量的距离确定目标向量所在的聚类集合，从而更加高效的找到目标向量。

[0021] 在上述第一方面的一种可能的实现中，其中，在所述当前节点的各聚类集合中确定第一预设数量的边缘点，包括：

[0022] 根据所述第一预设数量以及所述当前节点内各聚类集合的特征向量的数量，确定所述当前节点内各聚类集合的边缘点的第二预设数量；

[0023] 从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点。

[0024] 可以理解，本申请实施例中，根据第一预设数量以及当前节点内各聚类集合的特征向量的数量，确定当前节点内各聚类集合的边缘点的第二预设数量，从当前节点内各聚类集合中确定对应的第二预设数量的边缘点，可以使得在当前节点中获得的各聚类集合的边缘点的分布更加的合理。

[0025] 在上述第一方面的一种可能的实现中，其中，从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点，包括：

[0026] 确定所述当前节点内，每个聚类集合的相邻聚类集合；

[0027] 对于对应聚类集合的各特征向量，获取所述各特征向量与所述对应聚类集合的代表点之间的第一距离，并获取所述各特征向量与所述相邻聚类集合的代表点之间的第二距离；

[0028] 根据所述第一距离和所述第二距离，确定对应的第一序列，将所述第一序列中的第二预设数量个特征向量作为所述对应聚类集合的边缘点。

[0029] 可以理解，确定当前节点内，每个聚类集合的相邻聚类集合，对于对应聚类集合的各特征向量，获取各特征向量与对应聚类集合的代表点之间的第一距离，第一距离代表着各特征向量与对应聚类集合中中心点的距离远近，并获取各特征向量与相邻聚类集合的代表点之间的第二距离，第二距离代表着各特征向量与对应相邻聚类集合中中心点的距离远近，因而根据第一距离以及第二距离，确定对应聚类集合的边缘点更加合理、有效。

[0030] 在上述第一方面的一种可能的实现中，根据所述第一距离和所述第二距离，确定第一序列，将所述第一序列中的第二预设数量个特征向量作为所述对应聚类集合的边缘点，包括：

[0031] 获取所述第一距离和所述第二距离的距离之和；

[0032] 对所述各特征向量，根据所述距离之和由大至小排序，获取所述第一序列；

[0033] 将所述第一序列中的前第二预设数量个特征向量作为所述对应聚类集合的边缘点。

[0034] 可以理解的，根据第一距离和第二距离的距离之和从大到小进行排序得到第一序列，根据得到的第一序列确定距离之和较大的特征向量，即为距离第一聚类集合的代表点较远，且距离第二聚类集合的代表点也较远的特征向量，即该特征向量可能位于第一聚类集合的代表点指向相邻聚类集合的代表点的相反方向上的边缘点。当获取了第一聚类集合的各方向上的相邻聚类集合对应的第一序列，则可以获取第一聚类集合的各方向上的边缘点。如此，将第一聚类集合的各方向上的边缘点作为下一个节点的预设代表点，可以有效提高搜索精确度。

[0035] 在上述第一方面的一种可能的实现中，根据所述第一距离和所述第二距离，确定第一序列，将所述第一序列中的第二预设数量个特征向量作为所述对应聚类集合的边缘点，包括：

[0036] 获取所述第二距离和所述第一距离的差值与所述第一距离的比值；

[0037] 对所述各特征向量，根据所述比值由小至大排序，获取所述第一序列；

[0038] 将所述第一序列中的前第二预设数量个特征向量作为所述对应聚类集合的边缘点。

[0039] 可以理解的，本申请实施例中，获取第二距离和第一距离的差值与第一距离的比值，当比值越小表明该特征向量距离中心点越远，根据比值由小至大排序，将序列中的前第二预设数量个距离各特征向量所在聚类集合作为对应聚类集合的边缘点较为合理。

[0040] 在上述第一方面的一种可能的实现中，其中，从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点，包括：

[0041] 确定对应聚类集合中各特征向量与所述对应集合的代表点之间的第一距离；

[0042] 根据所述第一距离，从所述各特征向量中，确定第一序列，将所述第一序列中的第二预设数量个特征向量作为所述第一聚类集合的边缘点。

[0043] 可以理解的，本申请实施例中，确定当前节点的对应聚类集合中各特征向量与对应集合的代表点之间的第一距离，根据第一距离，当第一距离越大，表明特征向量距离中心点越远，则将第一序列中的距离较大的特征向量作为第一聚类集合的边缘点方便、合理。

[0044] 在上述第一方面的一种可能的实现中，其中，从所述当前节点内各聚类集合中确定对应的所述第二预设数量的边缘点，包括：

[0045] 确定所述当前节点内，每个聚类集合的相邻聚类集合；

[0046] 确定对应聚类集合中各特征向量与所述对应集合的相邻聚类集合的代表点之间的第二距离；

[0047] 根据所述第二距离，从所述各特征向量中，确定第一序列，将所述第一序列中的第二预设数量个特征向量作为所述第一聚类集合的边缘点。

[0048] 可以理解的，本申请实施例中，确定当前节点内，每个聚类集合的相邻聚类集合，确定对应聚类集合中各特征向量与对应集合的相邻聚类集合的代表点之间的第二距离，第二距离越小表明该特征向量距离相邻聚类集合越近，就是越接近该第一聚类集合的边缘，所以根据第二距离可以更加精确地确定第一聚类集合的边缘点。

[0049] 在上述第一方面的一种可能的实现中，所述距离包括欧式距离、内积距离和汉明距离。

[0050] 可以理解，本申请实施例中距离包括欧式距离、内积距离和汉明只是距离说明，也可以用其他任意可实施的距离表示。

[0051] 第二方面，本申请实施例提供了一种索引构建装置，包括：

[0052] 第一确定单元，用于确定目标向量以及所述目标向量在当前节点内所在的第一聚类集合，所述第一聚类集合具有第一代表点；

[0053] 第二确定单元，用于确定所述目标向量在下一节点内所在的第二聚类集合，所述第二聚类集合具有对应的代表点，所述第二聚类集合对应的预设代表点为所述第一节点内其中一个聚类集合的边缘点，其中，所述第一节点内的特征向量与所述第二节点内的特征向量相同；

[0054] 关联单元，分别在各所述节点中，建立所述目标向量与所述目标向量所在的第一聚类集合之间的索引关联。

[0055] 第三方面，本申请实施例提供了一种向量搜索方法，包括：

[0056] 获取查询向量；

[0057] 获取各节点中各聚类集合的代表点与查询向量的第三距离；

[0058] 根据所述第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离；

[0059] 根据所述第四距离，确定出所述查询向量在所述各节点中对应的第一目标向量；

[0060] 根据所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量。

[0061] 可以理解的，本申请的实施例中，在进行检索的过程中，可以将查询向量与各节点中的各聚类集合对应的代表点进行比较，获取各节点中各聚类集合的代表点与查询向量的第三距离，根据第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离，能够考虑到了距离中心点较远的边缘点与查询向量之间的距离，避免出现现有技术中的由于目标向量为边缘点，而现有技术中心只考虑了与中心点的距离，导致未找到更加精确的目标向量的问题。再者，根据第四距离，确定出查询向量在各节点中对应的第一目标向量，按照所述第四距离，从查询向

量在各节点中对应的第一目标向量中，能够更加全面的获得查询向量的目标向量，再基于第一目标向量获得查询向量对应的第二目标向量即最终确定的目标向量，得到的目标向量更加准确。

[0062] 在上述第三方面的一种可能的实现中，还包括，其中，按照所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量，包括：

[0063] 根据所述第四距离，对所述查询向量在所述各节点中对应的第一目标向量，按照所述第四距离由小至大排序，确定前设定次序的目标向量为所述查询向量对应的第二目标向量。

[0064] 可以理解，将查询向量在各节点中对应的第一目标向量查的距离确定前设定次序的目标向量作为查询向量对应的第二目标向量即最终目标向量，可以确保目标向量的准确性。

[0065] 第四方面，本申请实施例提供一种搜索装置，应用于数据系统，包括：

[0066] 第一获取单元，用于获取查询向量；

[0067] 第二获取单元，用于获取各节点中各聚类集合的代表点与查询向量的第三距离；

[0068] 第一确定单元，用于根据所述第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第二距离；

[0069] 第二确定单元，用于根据所述第四距离，确定出所述查询向量在所述各节点中对应的第一目标向量；

[0070] 第三确定单元，用于根据所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量。

[0071] 第五方面，本申请实施例提供一种检索系统，包括上述的索引构建装置，和/或上述的搜索装置。

[0072] 第六方面，本申请实施例提供一种索引结构，包括多个子索引结构，每个子索引结构包括多个代表点项和多个倒排文件项；所述代表点项包括数据系统中各个聚类集合所对应的各代表点，所述倒排文件项包括所述各代表点所对应的各倒排文件；

[0073] 所述每个倒排文件包括所述每个倒排文件对应的代表点所对应的聚类集合中的各个特征向量。

[0074] 第七方面，本申请实施例提供一种电子设备，包括：存储器，用于存储由电子设备的一个或多个处理器执行的指令，以及处理器，是所述电子设备的所述一个或多个处理器之一，用于执行权利要求上述第一方面任一项所述的索引构建方法或者上述第三方面任一项所述的向量搜索方法。

[0075] 第八方面，本申请实施例提供一种可读介质，包括，所述可读介质上存储有指令，该指令在电子设备上执行时使机器执行权利要求上述第一方面任一项所述的索引构建方法或者上述第三方面任一项所述的向量搜索方法。

[0076] 第九方面，本申请实施例提供一种计算机程序产品，包括指令，所述指令用于实现权利要求上述第一方面任一项所述的索引构建方法或者上述第三方面任一项所述的向量搜索方法。

[0077] 基于上述方案，本申请具有如下有益效果：

[0078] 本申请提供的索引构建方法，提供了精确选取当前节点的各聚类集合的边缘点的多种方式，且可以实现选取当前节点的第一聚类集合的各方向上的边缘点，并将当前节点的第一预设数量的边缘点作为下一节点的预设代表点，根据下一节点的预设代表点，获取所述下一节点的各聚类集合，分别在各节点中，建立目标向量与目标向量所在的第一聚类集合之间的索引关联。使得在进行检索的过程中，可以将查询向量与各节点的各聚类集合对应的代表点的距离均进行比较，从而根据比较结果确定目标代表点。如此，能够考虑到了距离中心点较远的边缘点与查询向量之间的距离，避免出现现有技术中的由于目标向量为边缘点，而现有技术中心只考虑了查询向量与中心点的距离，导致未找到更加精确的目标向量的问题，而遗漏掉其他与查询向量距离较远的第一代表点所属的聚类集合中存在的目标向量，有效提高搜索精确度。

附图说明

[0079] 图1a根据本申请的一些实施例，示出了一种图片数据库S的示意图；

[0080] 图1b根据本申请的一些实施例，示出了一种对图片数据库S进行倒排索引构建的示意图；

[0081] 图2根据本申请的一些实施例，示出了一种各节点对应的数据库的示意图；

[0082] 图3a根据本申请的一些实施例，示出了一种第一节点中各聚类集合的示意图；

[0083] 图3b根据本申请的一些实施例，示出了一种第二节点中各聚类集合的示意图；

[0084] 图3c根据本申请的一些实施例，示出了一种第三节点中各聚类集合的示意图；

[0085] 图3d示出根据本申请的一些实施例，示出了一种数据库S的索引结构示意图；

[0086] 图4根据本申请的一些实施例，示出了一种确定聚类集合S1的相邻聚类集合的示意图；

[0087] 图5a根据本申请的一些实施例，示出了一种获取第一聚类集合S1的边缘点的示意图；

[0088] 图5b根据本申请的一些实施例，示出了一种获取第一聚类集合S1的边缘点的示意图；

[0089] 图5c根据本申请的一些实施例，示出了一种获取第一聚类集合S1的边缘点的示意图；

[0090] 图5d根据本申请的一些实施例，示出了一种获取第一聚类集合S1的边缘点的示意图；

[0091] 图6a根据本申请的一些实施例，示出了一种获取查询向量与第一节点中各代表点的示意图；

[0092] 图6b根据本申请的一些实施例，示出了一种获取查询向量与第二节点中各代表点的示意图；

[0093] 图6c根据本申请的一些实施例，示出了一种获取查询向量与第三节点中各代表点的示意图；

[0094] 图7a根据本申请的一些实施例，示出了一种查询向量在第一节点对应的目标向量的示意图；

[0095] 图7b根据本申请的一些实施例，示出了一种查询向量在第二节点对应的目标向量

的示意图；

[0096] 图7c根据本申请的一些实施例，示出了一种查询向量在第三节点对应的目标向量的示意图；

[0097] 图8根据本申请的一些实施例，示出了一种索引构建方法的流程示意图；

[0098] 图9根据本申请的一些实施例，示出了一种第一节点和第二节点对应的的子索引结构；

[0099] 图10根据本申请的一些实施例，示出了一种搜索方法的流程示意图；

[0100] 图11根据本申请的一些实施例，示出了一种索引构建装置的示意图；

[0101] 图12根据本申请的一些实施例，示出了一种搜索装置的示意图；

[0102] 图13根据本申请的一些实施例，示出了一种检索系统的示意图；

[0103] 图14根据本申请的一些实施例，示出了一种电子设备的框图。

具体实施方式

[0104] 本申请的说明性实施例包括但不限于一种索引构建方法、装置、数据系统及搜索方法。

[0105] 如前所述，目前数据库中采用构建倒排索引的方式获得的检索结果的精确度会较低。

[0106] 例如，图1a中为一种检索系统中图片数据库S的示意图，数据库S中包括各图片对应的各特征向量。图1b中为对图片数据库S进行倒排索引构建的示意图。如图1b所示，首先通过聚类处理将数据库S中各特征向量分配至四个聚类集合中，分别为聚类集合S1、聚类集合S2、聚类集合S3和聚类集合S4，其中，聚类集合S1具有对应的代表点C1，聚类集合S2具有对应的代表点C2，聚类集合S3具有对应的代表点C3，和聚类集合S4具有对应的代表点C4。

[0107] 可以理解，本申请实施例中提及的代表点可以是聚类集合的中心点，即与各个点的距离之间的差异在预设范围内的点；也可以是基于其他的规则制定的点。

[0108] 在基于上述数据库S对查询图片所对应的查询向量A进行向量检索时，检索系统会确定出每个代表点与查询向量A之间的距离，例如，如图1b所示，检索系统会确定出代表点C1与查询向量A之间的距离为d1，代表点C2与查询向量A之间的距离为d2，代表点C3与查询向量A之间的距离为d3，代表点C4与查询向量A之间的距离为d4；假设d4＞d3＞d1＞d2时，系统会确定出四个代表点中与查询向量A之间的距离最近的代表点为代表点C2，然后对代表点C2所在的聚类集合S2中的每个特征向量进行搜索，可以理解，对每个特征向量进行搜索即为获取每个特征向量与查询向量A之间的距离。此后，将聚类集合S2中与查询向量A较近的或处于设定距离范围内的特征向量Y1和特征向量Y2作为查询向量A的目标向量。然后将目标向量对应的原图片数据输出至客户端。

[0109] 但是，如图1b所示，实质上聚类集合S1中的特征向量X1却是数据库S中距离查询向量A最近的特征向量，但由于聚类集合S1的代表点C1不是距离查询向量最近的，因此，系统并未对聚类集合S1中的特征向量进行搜索，使得未获取到准确的查询向量A的目标向量，导致漏掉了最确切的检索结果，影响检索的精确度。

[0110] 为了解决上述问题，本申请实施例提供了一种索引的构建方法，具体包括：获取待创建索引的数据库中各数据所对应的特征向量；将获取的各特征向量冗余(或称为重复)放

置在预设数量的节点中，其中，每个节点内的特征向量相同。

[0111] 可以理解的，在检索系统为分布式系统，即具有多个处理器和多个对应的存储器系统时，上述预设数量的节点中的每个节点同时具有计算和存储能力，是计算资源和存储资源的一个集合。因此，不同节点的检索可以并行处理。在一些实施例中，当检索系统只有单处理器和单存储器时，上述节点可以指单个存储器中进程、线程或协程以及单个存储器上不同存储区域的集合。

[0112] 对于第一节点，预设若干代表点，并基于预设的代表点对第一节点内的特征向量进行聚类处理，将第一节点内的特征向量分配至对应的聚类集合中，每个聚类集合对应的代表点；然后根据第二节点的聚类集合的第一预设数量，从第一节点的各聚类集合中确定出第一预设数量的边缘点，作为第二节点的各聚类集合的预设代表点，基于第二节点的各聚类集合的预设代表点对第二节点内的特征向量进行聚类处理，将第二节点内的特征向量分配至第二节点内对应的聚类集合中，第二节点内对应的各聚类集合具有对应的代表点。

[0113] 后续节点内的特征向量均参照第二节点的聚类方式，即根据当前节点的聚类集合的预设数量，从前一节点的各聚类集合中，获取预设数量的边缘点作为当前节点的各聚类集合的预设代表点，基于当前节点的各聚类集合的预设代表点对当前节点内的特征向量进行聚类处理得到对应的各聚类集合，其中，得到的各聚类集合具有对应的代表点。

[0114] 若已对预设数量的节点均进行了聚类处理，则对于各目标向量，确定各目标向量在各节点内所在的聚类集合；将目标向量与目标向量在各节点内所在的聚类集合对应的代表点建立索引关联。即在各节点均建立一个子索引结构。例如，假设目标向量在第一节点被分配至第一节点内的第一聚类集合中，则在第一节点的子索引结构中，建立第一节点与第一聚类集合的代表点之间的索引关联；假设目标向量在第二节点被分配至第二节点内的第二聚类集合中，则在第二节点的子索引结构中，建立第二节点与第二聚类集合的代表点之间的索引关联。

[0115] 可以理解，上述对各节点内特征向量进行聚类处理的方式均可以为K-means等聚类方式。其中，K-means聚类方式可以为获取数据库中每个特征向量与各聚类集合对应的代表点的距离；将各特征向量分别分配至距离最近的代表点所对应的聚类集合中。

[0116] 下面以对图1a中的数据库S进行索引构建，说明本申请实施例中的索引构建方法。

[0117] 例如，将图1a中的数据库S中对应的特征向量分别重复放置在三个节点中，得到如图2所示的第一节点、第二节点和第三节点。可以理解的，第一节点、第二节点和第三节点中存储的数据库S均相同。

[0118] 对于第一节点，对第一节点内的各特征向量进行聚类处理，获取如图3a所示的第一节点内的4个聚类集合S1、S2、S3、S4以及这4个聚类集合分别对应的代表点C1、C2、C3、C4。在第一节点的子索引结构中，将第一节点的代表点C1、C2、C3、C4分别与各代表点所属聚类集合中的各特征向量建立第一节点对应的索引关联。

[0119] 对于第二节点，假设第二节点对应的聚类集合的第一预设数量是4，则可以在第一节点中确定出4个边缘点。例如，如图3a中示出了，在第一节点中选出的边缘点为聚类集合S1中的边缘点C1’、聚类集合S2中的边缘点C2’、聚类集合S3中的边缘点C3’、聚类集合S4中的边缘点C4’。其中，选取方法将在下文阐述，在此不再赘述。然后如图3b所示，将上述在第一节点中选出的边缘点C1’、C2’、C3’、C4’作为第二节点的预设代表点，对第二节点的所有

特征向量进行聚类处理，得到第二节点对应的各聚类集合S1’、S2’、S3’和S4’。在第二节点的子索引结构中，将第二节点的代表点C1’、C2’、C3’、C4’分别与各代表点所属聚类集合S1’、S2’、S3’、S4’中的各特征向量建立第二节点对应的索引关联。

[0120] 可以理解，在一些实施例中，根据任意节点的预设代表点对节点内特征向量进行聚类后获得对应的聚类集合后，可以重新计算各聚类集合的代表点。例如，将各聚类集合中各特征向量的算术平均值作为该聚类集合的代表点。可以理解，重新计算的各聚类集合的代表点与初始预设的代表点一般为不一致的。但为了方便描述，本申请的实施例中以初始预设代表点近似替代聚类后的代表点为例对本申请实施例中的技术方案进行介绍。

[0121] 对于第三节点，假设第三节点的集合预设数量是3，在第二节点的各聚类集合中选出总个数为3的边缘点，并将这3个边缘点作为第三节点的预设代表点，假设在第二节点中选出如图3c所示的预设数量的边缘点，具体地，选出的边缘点为聚类集合S1’中的边缘点C1”、聚类集合S2’中的边缘点C2”，聚类集合S3’中的边缘点C3”，则第三节点的预设代表点为第二节点的边缘点C1”、C2”、C3”，选取方法将在下文阐述，在此不再赘述。根据预设代表点，对第三节点的所有特征向量进行聚类处理，得到如图3c所示的第三节点对应的3个聚类集合S1”、S2”、S3”，以及这3个聚类集合对应的代表点C1”、C2”、C3”。在第三节点的子索引结构中，将第三节点的代表点C1”、C2”、C3”分别与各代表点所属聚类集合S1”、S2”、S3”中的各特征向量建立第三节点对应的索引关联。。

[0122] 图3d示出了上述数据库S的索引结构示意图，分别包括第一节点的子索引结构、第二节点的子索引结构和第三节点的子索引结构。

[0123] 如图3d所示，第一节点对应的子索引结构包括代表点项和倒排文件项。代表点项包括聚类集合S1具有对应的代表点C1，聚类集合S2具有对应的代表点C2，聚类集合S3具有对应的代表点C3，和聚类集合S4具有对应的代表点C4。

[0124] 聚类集合S1的对应的代表点C1具有对应的倒排文件D1，该倒排文件D1中包括对应的代表点C1所对应的聚类集合S1中的各个特征向量。

[0125] 聚类集合S2的对应的代表点C2具有对应的倒排文件D2，该倒排文件D2中包括对应的代表点C2所对应的聚类集合S2中的各个特征向量。

[0126] 聚类集合S3的对应的代表点C3具有对应的倒排文件D3，该倒排文件D3中包括对应的代表点C3所对应的聚类集合S3中的各个特征向量。

[0127] 聚类集合S4的对应的代表点C4具有对应的倒排文件D4，该倒排文件D4中包括对应的代表点C4所对应的聚类集合S4中的各个特征向量。

[0128] 其中，第二节点的子索引结构和第三节点的子索引结构与第一节点的子索引结构类似，下面不再赘述。

[0129] 可以理解，在第二节点的聚类集合的预设数量，即第一预设数量确定后，可以根据以下方法确定第一节点各聚类集合中所需要确定的边缘点的第二预设数量：

[0130] 第一种可实施的方案中，确定第一节点中各聚类集合的边缘点的第二预设数量的方法为：根据第一预设数量，按照第一节点中各聚类集合的半径，根据第一预设数量以及各个聚类集合的半径的比例确定每个聚类集合的边缘点的第二预设数量；可以理解，在一些实施例中，当计算出来的任一聚类集合的边缘点的预设数量不为整数时，可以根据取整规则获取最终的预设数量。例如，取整规则可以为采取将当前数量对应的数值的整数位加1，

舍去小数部分所获得的数值，作为最终的的预设数量所对应的数值。

[0131] 例如，第一预设数量为4，即边缘点预设总数量为4，如图3a所示的第一节点对应的聚类集合中，假设聚类集合S1半径大小为16、聚类集合S2半径大小为14、聚类集合S3半径大小为10以及聚类集合S4半径大小为17，则聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的数量比例为16:14:10:17；根据边缘点的预设总数量和上述边缘点的数量比例为16:14:10:17可以确定出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的数量为1.1，0.9，0.7和1.7。根据上述取整规则，可以得出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的第二预设数量分别为2，1，1和2。

[0132] 可以理解，本申请实施例中，通过将各节点中预设数量的边缘点作为下一节点的预设代表点的方式可以使得各节点的聚类尽可能不同，如此，可以有效提高搜索精度。

[0133] 其次，上述根据第一预设数量以及各个聚类集合的半径的比例确定每个聚类集合的边缘点的预设数量的方式可以使得边缘点的分布更加均匀，避免出现有些聚类集合半径很小，但还确定了数量较多的边缘点的情况的发生。

[0134] 第二种可实施的方案中，确定第一节点中各聚类集合的边缘点的第二预设数量的方法为：根据第一预设数量，按照第一节点中各聚类集合内半径从大到小排序，从每个聚类集合中选取设定个数的边缘点，直到选完预设第一数量个。

[0135] 例如，第一预设数量为4，即边缘点预设总数量为4，如图3a所示的第一节点对应的聚类集合中，假设聚类集合S1半径大小为16、聚类集合S2半径大小为14、聚类集合S3半径大小为10以及聚类集合S4半径大小为17，则按照各个聚类集合半径大小从大到小排列得到的顺序为聚类集合S4-聚类集合S1-聚类集合S2-聚类集合S3；当每个聚类集合中选取的边缘点的设定个数为2个，按照聚类集合S4-聚类集合S1-聚类集合S2-聚类集合S3的先后顺序，聚类集合S4和聚类集合S1中需要确定的边缘点的各第二预设数量均为两个。因为此时已达到边缘点的预设总数量。后续聚类集合S2和聚类集合S3将不再选取边缘点，即排在聚类集合S1后面的聚类集合S2和聚类集合S3的个数为0个。

[0136] 可以理解，上述根据第一预设数量，按照各个聚类集合的半径从大到小排序，从每个聚类集合中选取设定个数的边缘点的方式可以使得在半径较小聚类集合中确定数量较少的边缘点，而在其他半径较大的聚类集合确定较多的边缘点，可以使得确定的边缘点能够充分代表对应的聚类集合，有效提高搜索精度。

[0137] 第三种可实施的方式中，确定第一节点中各聚类集合的边缘点的第二预设数量的方法为：确定第一节点中各聚类集合的代表点与其最相近的设定数量的特征向量之间的距离之和，根据各聚类集合所对应的距离之和的比例以及第一预设数量确定各聚类集合的边缘点的第二预设数量。

[0138] 可以理解，上述选取的与各聚类集合的代表点相近的特征向量的设定数量可以为根据实际需求设置。

[0139] 可以理解，当聚类集合的代表点与其最相近的设定数量的特征向量之间的距离之和较大，则能在一定程度上证明该聚类集合中的特征向量大多都距离中心点较远，因此可能存在较多的边缘点，此时设置该聚类集合中应该获取的第二预设数量较多，能够使得该较大的聚类集合中确定的边缘点能够充分代表该聚类集合，有效提高搜索精度。

[0140] 例如，第一预设数量为4，即边缘点预设总数量为4，如图3a所示的第一节点对应的聚类集合中，假设聚类集合S1中代表点C1与其最相近的设定数量的特征向量之间的距离之和为16、聚类集合S2中代表点C2与其最相近的设定数量的特征向量之间的距离之和为14、聚类集合S3代表点C3与其最相近的设定数量的特征向量之间的距离之和为10，聚类集合S4代表点C4与其最相近的设定数量的特征向量之间的距离之和为17。则聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的数量比例为16:14:10:17；根据边缘点的预设总数量4和上述边缘点的数量比例为16:14:10:17可以确定出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的数量为1.1，0.9，0.7和1.7。根据上述取整规则，可以得出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的第二预设数量分别为2，1，1和2。

[0141] 可以理解，确定其他节点内的各聚类集合所需要确定的边缘点的第二预设数量的方式均可以采用上述方式。

[0142] 在基于上述方式确定了在各聚类集合所需要获取的边缘点的第二预设数量后，则进一步可以在对应聚类集合内确定对应的第二预设数量的边缘点。下面以在第一节点内的第一聚类集合中确定第二预设数量的边缘点的方式为例，对本申请实施例中提供的在任意聚类集合中确定第二预设数量的边缘点的方式进行介绍：

[0143] 第一种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：获取第一聚类集合的代表点与该第一聚类集合内的所有特征向量的第一距离；按照第一距离将各特征向量由远到近排序，得到第一聚类集合对应的第一序列；确定第一序列中前第二预设数量个特征向量为该聚类集合内的边缘点。

[0144] 可以理解，本申请实施例中，可以根据第一聚类集合中各特征向量与代表点之间的距离确定边缘点，便于能够将距离代表点较远的特征向量作为边缘点，使得边缘点的选取更加精确。

[0145] 且本申请实施例中，根据第一聚类集合中各特征向量与代表点之间的距离，即根据各特征向量与代表点的远近情况，将各特征向量从远到近排序，可以使得根据设定数量选取对应的边缘点时更加方便。例如，边缘点的设定数量为五个，则可以直接将排序即第一序列中的前五个特征向量作为边缘点。

[0146] 第二种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：获取第一聚类集合的代表点与该第一聚类集合内所有特征向量的第一距离；将所述第一距离大于设定距离的特征向量作为第一序列中的特征向量，然后将第一序列中的第二预设数量的特征向量作为边缘点，或者将第一序列中的各特征向量按照第一距离进行由远至近的排序，将排序中的前第二预设数量的特征向量作为边缘点。

[0147] 可以理解，本申请实施例中，可以根据第一聚类集合中各特征向量与代表点之间的距离大于设定距离的特征向量获得第一序列，即第一序列中的特征向量是距离代表点较远的且大于设定距离的向量，且可以使得选取相邻聚类集合的方式更加规范且简单。例如，可以只需在选取边缘点的算法中设置相应的距离阈值参数即可实现边缘点的获取。

[0148] 第三种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：确定除第一聚类集合之外的其他聚类集合中，与第一聚类集合邻近的至少一个相邻聚类集合；获取第一聚类集合中每个特征向量分别与每个相邻聚类集合的代表点的距离；

根据该距离，对各特征向量由近至远排序，获取每个相邻聚类集合对应的第一序列；将第一序列中的前设定数量个特征向量作为边缘点。

[0149] 其中，获取第一聚类集合的邻近的至少一个相邻聚类集合的方式可以为：将上述多个聚类集合中除第一聚类集合之外的各聚类集合，按照各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行由近至远的排序，获取该第一聚类集合所对应的第二序列；将所述第二序列中的前第二预设数量个聚类集合作为与第一聚类集合邻近的相邻聚类集合。

[0150] 下面以前述图3a中的第一聚类集合S1为例，介绍获取第一聚类集合S1中的边缘点的方式：

[0151] 首先可以获取第一聚类集合S1的邻近的至少一个相邻聚类集合。其中获取第一聚类集合S1的邻近的至少一个相邻聚类集合的方式可以为：获取与除聚类集合S1之外的其他三个聚类集合的代表点与第一聚类集合S1的代表点C1之间的距离。例如，如图4所示，可以获取到聚类集合S2的代表点C2与第一聚类集合S1的代表点C1与的距离为d5，聚类集合S3的代表点C3与第一聚类集合S1的代表点C1与的距离为d6，聚类集合S4的代表点C4与第一聚类集合S1的代表点C1与的距离为d7，然后根据上述三个聚类集合各自的代表点与第一聚类集合S1的代表点C1之间的距离由近到远，对上述三个聚类集合进行排序，获取代表点C1对应的第二序列；假设d7＜d5＜d6时，则代表点C1对应的第二序列为聚类集合S4-聚类集合S2-聚类集合S3；此时，将代表点C1与所述第二序列中的前第二预设数量个聚类集合建立索引关联。假设第二预设数量为2个，则可以将聚类集合S2、聚类集合S4作为第一聚类集合S1的邻近的相邻聚类集合。

[0152] 然后，获取第一聚类集合S1中每个特征向量与相邻聚类集合S2的代表点C2的距离，以及第一聚类集合S1中每个特征向量与相邻聚类集合S4的代表点C2的距离。如图5a中所示，对于特征向量X1，获取第一聚类集合S1中特征向量X1与相邻聚类集合S2的代表点C2的距离d2_1，对于其他特征向量与代表点C2的距离在此不再赘述；根据各特征向量与代表点C2的距离，对各特征向量由近至远排序，假设根据距离，将第一聚类集合S1中15个特征向量由远至近排序得到的第一序列为X1-X3-X5-.....-X15。假设预设的根据每个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个，则可以将相邻聚类集合S2对应的第一序列中的第一个特征向量，即特征向量X1作为边缘点。同时，获取第一聚类集合S1中每个特征向量与相邻聚类集合S4的代表点C4的距离，如图5a中所示，对于特征向量X1，获取第一聚类集合S1中特征向量X1与相邻聚类集合S4的代表点C4的距离d4_1，对于其他特征向量与代表点C4的距离在此不再赘述；根据各特征向量与代表点C4的距离，对各特征向量由近至远排序，假设根据距离，将第一聚类集合S1中15个特征向量由近至远排序得到的第一序列为X7-X8-X1-.....-X14。假设预设的根据每个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个，则可以将第四聚类集合S4对应的第一序列中的第一个特征向量，即特征向量X7作为边缘点。

[0153] 可以理解的，当第二距离越小就说明该特征向量越接近相邻聚类集合，也就是越接近该第一聚类集合的边缘。因此，上述通过第一聚类结合中各特征向量与各相邻聚类集合的距离获取边缘点的方式，便于获取第一聚类集合在指向各相邻聚类集合的方向上的边缘点。如此，便于在下一节点内进行检索时，能够根据边缘点与查询向量的距离确定出与查

询向量最为接近的代表点，并对该代表点对应的聚类集合进行搜索，对于与各聚类集合的代表点的距离较远，而与边缘点即边缘点的距离较近的查询向量，在查询的过程中能够增大目标向量被查找到的概率，有效提升搜索精度。

[0154] 可以理解，上述通过按照各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行排序来获取第一聚类集合的邻近聚类集合，即相邻聚类集合的方式只是示例性说明。可实施的，本申请实施例获取第一聚类集合的邻近聚类集合即相邻聚类集合的方式也可以为其他方式。

[0155] 其中，获取相邻聚类集合的方式还可以为：将上述多个聚类集合中除第一聚类集合之外的各聚类集合的代表点与第一聚类集合的代表点之间的距离小于设定距离的聚类集合组成第二序列中的聚类集合；然后将第二序列中的任意第二预设数量的聚类集合作为相邻聚类集合，或者将第二序列中的各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行由近至远的排序，将排序中的前第二预设数量的聚类集合作为相邻聚类集合。

[0156] 例如，如图4所示，确定与聚类集合S1邻近的相邻聚类集合的方式可以为：首先获取除聚类集合S1之外的其他聚类集合的代表点与聚类集合S1的代表点之间的距离。例如，聚类集合S2的代表点C2与第一聚类集合S1的代表点C1与的距离为d5，聚类集合S3的代表点C3与第一聚类集合S1的代表点C1与的距离为d6，聚类集合S4的代表点C4与第一聚类集合S1的代表点C1与的距离为d7。然后将d5、d6以及d7与设定距离D进行比较，假设d5、d7小于设定距离D，则聚类集合S2、聚类集合S4组成第二序列中的聚类集合；然后将第二序列中的聚类集合S2、聚类集合S4中的代表点与该聚类集合S1的代表点之间的距离d5、d7进行由近至远的排序，假设d7小于d5，则排序为聚类集合S4-聚类集合S2。若相邻聚类集合的第二预设数量为一个，则将聚类集合S4作为相邻聚类集合。

[0157] 再例如，获取相邻聚类集合的方式还可以为：估算除第一聚类集合的代表点本身所在的第一聚类集合之外的各聚类集合的近似半径距离，获取各聚类集合的代表点与第一聚类集合的代表点的之间的距离与近似半径距离的差值，根据差值，将各聚类集合进行由近至远排序获取第二序列，将第二序列中的前第二预设数量个聚类集合作为相邻聚类集合。

[0158] 可以理解，在一些实施例中，上述估算除目标特征向量本身所在的第一聚类集合之外的各聚类集合的近似半径距离的方式可以为：获取各聚类集合中所有特征向量与各聚类集合中的代表点之间的距离，将所有特征向量与代表点之间的最大距离作为各聚类集合的近似半径距离。

[0159] 例如，如图5b中所示，确定与聚类集合S1邻近的相邻聚类集合的方式可以为：首先，确定除聚类集合S1之外的各聚类集合S2、S3以及S4，将聚类集合S2中各特征向量中与聚类集合S2的代表点C2的最远距离作为聚类集合S2的近似半径距离，假设最远距离为特征向量Y3与聚类集合S2的代表点C2之间的距离r1，则将r1作为聚类集合S2的近似半径距离；将聚类集合S3中各特征向量中与第一中心点C3的最远距离作为聚类集合S2的近似半径距离，假设最远距离为特征向量Z1与聚类集合S3的代表点C3之间的距离r2，则将r2作为聚类集合S3的近似半径距离；将聚类集合S4中各特征向量中与第一中心点C4的最远距离作为聚类集合S4的近似半径距离，假设最远距离为特征向量W5与聚类集合S4的代表点C4之间的距离r3，则将r3作为聚类集合S4的近似半径距离r3。

[0160] 然后，确定聚类集合S2的代表点C2与聚类集合S1的代表点C1与的距离d5与聚类集合S2的半径距离r1的差值为d5-r1，聚类集合S3的代表点C3与聚类集合S1的代表点C1与的距离d6与聚类集合S3的半径距离r2的差值为d6-r2，聚类集合S4的代表点C4与聚类集合S1的代表点C1与的距离d7与聚类集合S4的半径距离r3的差值为d7-r3，根据差值，将聚类集合S2、S3以及S4进行由近至远排序获取第二序列，假设d7-r3<d5-r1<d6-r2，则第二序列为聚类集合S4-聚类集合S2-聚类集合S3，将第二序列中的前第二预设数量个聚类集合作为相邻聚类集合，假设将第二序列中的前1个数量的聚类集合作为相邻聚类集合，则相邻聚类集合为聚类集合S4。

[0161] 在一些实施例中，估算各聚类集合的近似半径的方式也可以为其他任意可实施的方式，例如通过神经网络模型或者相关算法估算等。

[0162] 可以理解，任一聚类集合的代表点与第一聚类集合的代表点之间的距离与近似半径的差值可以近似代表该聚类集合中的边缘特征向量与第一聚类集合的代表点之间。当边缘特征向量距离第一聚类集合的代表点之间的距离较近，则可能存在该边缘特征向量所在聚类集合为第一聚类集合的邻近集合。

[0163] 第四种可实施的方案中，确定符合预设条件的至少一个边缘点的方式可以为：确定除第一聚类集合之外的其他聚类集合中，与所述第一聚类集合邻近的至少一个相邻聚类集合；获取第一聚类集合中每个特征向量分别与所述第一聚类集合的代表点之间的第一距离以及与所述每个相邻聚类集合的代表点之间的第二距离，并根据第一距离和第二距离的距离之和；对各特征向量，按照距离之和由远至近排序，获取每个相邻聚类集合对应的所述第一序列；将第一序列中的前设定数量个特征向量作为边缘点。可以理解，第四种方案中获取第一聚类集合邻近的至少一个相邻聚类集合的方式与第三种方案中阐述的获取相邻聚类集合的方式相同，此处不在赘述。

[0164] 可以理解的，根据第一距离和第二距离的距离之和从大到小进行排序得到第一序列，得到的第一序列中距离之和较大的特征向量，即为距离第一聚类集合的代表点较远，且距离相邻聚类集合的代表点也较远的特征向量，即该特征向量可能位于代表点的代表点指向边缘点的相反方向上的边缘点。当获取了第一聚类集合的各方向上的相邻聚类集合对应的第一序列，则可以获取第一聚类集合的各方向上的边缘点。如此，将第一聚类集合的各方向上的边缘点作为边缘点，可以有效提高搜索精确度。

[0165] 下面以获取图3a中第一聚类集合S1的至少一个边缘点为例，说明本申请实施例中第四种实施方案中获取边缘点的方式：

[0166] 如图5c中所示，假设第一聚类集合S1邻近的相邻聚类集合为前述图4中所述的聚类集合S2和聚类集合S4。根据上述确定的第一聚类集合S1的邻近的相邻聚类集合S2，获取第一聚类集合S1中每个特征向量分别与该第一聚类集合S1的代表点C1之间的第一距离，以及与相邻聚类集合S2的代表点C2的第二距离的距离之和。

[0167] 例如，如图5c中所示，第一聚类集合S1中的特征向量X1与代表点C1之间的第一距离d1_1以及与代表点C2的第二距离d2_1的距离之和为d_s2，其他14个特征向量获得距离之和的方式相同，在此不再赘述；假设根据距离之和将第一聚类集合S1中15个特征向量由远至近排序得到的聚类集合S2对应的第一序列为X2-X3-X1-.....-X14，假设预设的根据每个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个，则可以将相邻聚类集

合S2对应的第一序列中的第一个特征向量,即特征向量X2作为边缘点。

[0168] 同时,根据上述确定的第一聚类集合S1的邻近的相邻聚类集合S4,获取第一聚类集合S1中每个特征向量分别与该第一聚类集合S1的代表点C1之间的第一距离以及与相邻聚类集合S4的代表点C4的第二距离的距离之和,例如,图4中,第一聚类集合S1中的特征向量X1与代表点C1的第一距离d1_1以及代表点C4的第二距离d4_1的距离之和d_s4,其他14个特征向量获得距离之和的方式相同,在此不再赘述;假设根据距离之和将第一聚类集合S1中15个特征向量由远至近排序得到的第一序列为X4-X1-X3-.....-X15,假设预设的根据每个邻近集合的获取的对应聚类集合的预设边缘点的数量为一个,则可以将第四聚类集合S4对应的第一序列中的第一个特征向量,即特征向量X4作为边缘点。

[0169] 第五种可实施的方案中,从第一聚类集合中确定第二预设数量的边缘点的方式可以为:确定除第一聚类集合之外的其他聚类集合中,与所述第一聚类集合邻近的至少一个相邻聚类集合;获取第一聚类集合中每个特征向量分别与所述第一聚类集合的代表点之间的第一距离以及与所述每个相邻聚类集合的代表点之间的第二距离。并根据第二距离和第一距离的差值与第一距离的比值,对第一聚类集合中各特征向量,按照比值从小到大排序,获取每个相邻聚类集合对应的第一序列;将第一序列中的前第二预设数量个特征向量作为边缘点。可以理解,第五种方案中获取该聚类集合邻近的至少一个相邻聚类集合的方式与第三种方案中阐述的获取相邻聚类集合的方式相同,此处不在赘述。

[0170] 可以理解的,当特征向量与相邻聚类集合的代表点之间的第二距离越小,,特征向量与第一聚类集合的代表点之间的第一距离越大,则证明特征向量距离第一聚类集合的代表点越远,即第二距离和第一距离的差值与第一距离的比值越小。因此,按照第二距离和第一距离的差值与第一距离的比值由小到大对特征向量进行排序,可以获得第一聚类集合的更为精确的边缘点。

[0171] 例如,下面以获取图3a中聚类集合S1的第二预设数量的边缘点为例,说明本申请实施例中第五种实施方案中获取边缘点的方式:

[0172] 假设聚类集合S1邻近的相邻聚类集合为前述图4中所述的聚类集合S2和聚类集合S4。根据上述确定的聚类集合S1的邻近的相邻聚类集合S2,获取聚类集合S1中每个特征向量分别与该聚类集合S1的代表点C1之间的第一距离,以及与相邻聚类集合S2的代表点C2的第二距离的距离之和。

[0173] 例如,如图5d中所示,聚类集合S1中的特征向量X1与代表点C2之间的第二距离d2_1以及与代表点C1的第一距离d1_1的距离差值与第一距离d1_1的比值为d_b2,其他14个特征向量获得距离之比的方式相同,在此不再赘述;假设根据比值由小至大将聚类集合S1中15个特征向量排序得到的聚类集合S2对应的第一序列为X8-X3-X1-.....-X14,假设预设的根据每个邻近集合的获取的对应聚类集合的相邻代表点的预设数量为一个,则可以将相邻聚类集合S2对应的第一序列中的第一个特征向量,即特征向量X8作为边缘点。

[0174] 同时,根据上述确定的聚类集合S1的邻近的相邻聚类集合S4,获取聚类集合S1中的特征向量X1与代表点C4之间的第二距离d4_1以及与代表点C1的第一距离d1_1的距离差值与第一距离d1_1的比值为d_b4,其他15个特征向量获得比值的方式相同,例如,图5d中,聚类集合S1中的特征向量X1与代表点C4的第二距离d4_1以及代表点C1的第一距离d1_1的距离差值与第一距离d1_1的比值为d_b4,其他14个特征向量获得比值的方式相同,在此不

再赘述;假设根据比值由小到大将聚类集合S1中15个特征向量排序得到的第一序列为X9-X1-X3-.....-X15,假设预设的根据每个邻近集合的获取的对应聚类集合的预设第二代表点的数量为一个,则可以将聚类集合S4对应的第一序列中的第一个特征向量,即特征向量X9作为边缘点。

[0175] 基于本申请的索引构建方法,可以使得在查询时,可以获取查询向量距离各节点内的聚类集合的代表点即中心点的距离,并可以获取查询向量与各节点的代表点之间的距离,并获取查询向量在各节点内对应的目标向量。然后可以将查询向量在各节点内对应的目标向量汇总进行对比,以实现搜索到与查询向量距离更近的,或更为精确的目标向量,能够有效提高搜索精度。

[0176] 且在索引构建时,第二节点与之后的节点均是以前一节点的边缘点作为当前节点的代表点进行聚类。如此,能够考虑到了距离中心点较远的边缘点与查询向量之间的距离,避免出现现有技术中的由于目标向量为边缘点,而现有技术中心只考虑了与中心点的距离,导致未找到更加精确的目标向量的问题。

[0177] 基于本申请的索引构建方法,相同的特征向量在不同的节点中对应的聚类集合以及聚类集合的代表点不同,而且以第一节点中各聚类集合的边缘点作为第二节点的代表点,这样在搜索的过程中可以增大目标向量被搜索到的概率,从而提高搜索的准确性。

[0178] 与上面的向量索引构建方法相应地,在针对这样的向量进行检索时,检索系统可以先获取查询向量;获取各节点中各聚类集合的代表点与查询向量的第三距离;根据所述第三距离,确定出各节点内的目标代表点,并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离;根据所述第四距离,确定出所述查询向量在所述各节点中对应的第一目标向量;按照所述第四距离,从所述查询向量在所述各节点中对应的第一目标向量中,确定所述查询向量对应的第二目标向量即最终确定的目标向量。

[0179] 例如,基于上述构建完成的如图3d所示的索引结构,检索系统可以先获取查询向量A,将查询向量A分发到第一节点、第二节点以及第三节点中。然后分别确定出第一节点中各第一聚类集合的代表点与查询向量A之间的第三距离、第二节点中各第二聚类集合的代表点与查询向量A之间的第三距离以及第三节点中各第三聚类集合的代表点与查询向量A之间的第三距离。例如图6a示出了获取查询向量A与第一节点中聚类集合S1的代表点C1、聚类集合S2的代表点C2、聚类集合S3的代表点C3以及聚类集合S4的代表点C4的第三距离分别为da1、da2、da3以及da4。例如图6b示出了获取查询向量A与第二节点中聚类集合S1'的代表点C1'、聚类集合S2'的代表点C2'、聚类集合S3'的代表点C3'以及聚类集合S4'的代表点C4'的第三距离分别为dm1、dm2、dm3以及dm4。例如图6c示出了获取查询向量A与第三节点中聚类集合S1"的代表点C"、聚类集合S2"的代表点C2"、聚类集合S3"的代表点C3"的第三距离分别为dk1、dk2以及dk3。

[0180] 根据3个节点中查询向量A与每个聚类集合中各代表点之间的第三距离,确定出各节点中与查询向量距离目标代表点,在一些实现方式中,将各节点中的各代表点根据与目标向量的第三距离进行去重和排序,在各节点中选择最近的前预设数量个代表点作为目标代表点,例如,在各节点中选择距离最近的一个代表点作为各节点的目标代表点,假设第一节点中与查询向量A距离最近的代表点为C1,第二节点中与查询向量A距离最近的代表点为

C1’，第三节点中与查询向量A距离最近的代表点为C2’，则将第一节点中与查询向量A距离最近的代表点C1，第二节点中与查询向量A距离最近的代表点C1’，第三节点中与查询向量A距离最近的代表点C2’为各节点对应的目标代表点。

[0181] 根据如图3d所示的索引关联，获取第一节点中代表点为C1所在聚类集合S1、第二节点中代表点为C1’所在聚类集合S1’、第三节点中代表点为C2”所在聚类集合S2”中各特征向量与查询向量的第四距离。

[0182] 在一些实施例中，可以将与查询向量A之间的第四距离满足小于设定值的特征向量作为第一目标向量。例如，将第一节点中聚类集合S1、第二节点中聚类集合S1’以及第三节点中聚类集合S2”中与查询向量A之间的距离在设定范围内的特征向量作为查询向量A在各节点中对应的目标向量，如图7a所示，假设第一节点中聚类集合S1中与查询向量A在设定范围内的特征向量为特征向量X1、特征向量X2且对应的第四距离分别为dn1、dn2，如图7b所示，第二节点中聚类集合S1’中与查询向量A在设定范围内的特征向量为特征向量X1’、特征向量X2’、特征向量X3’且对应的第四距离分别为dt1、dt2、dt3，如图7c所示，第三节点中聚类集合S2”中与查询向量A在设定范围内的特征向量为特征向量Y1’、特征向量Y2’且对应的第四距离分别为dw1、dw2。则第一节点对应的征向量X1、特征向量X2，第二节点对应的特征向量X1’、特征向量X2’、特征向量X3’以及第三节点对应的特征向量Y1’、特征向量Y2’中为查询向量对应的的第一目标向量。根据获取的第一目标向量与查询向量A对应的第四距离，可以在获取的第一目标向量中确定与查询向量A最近的前预设数量个第二目标向量。例如，预设的第二目标向量的数量为两个，假设dw1<dn1<dn2<dt1<dt2<dt3<dw2则根据第四距离确定第三节点中的Y1”、以及第一节点中的X1为第二目标向量。

[0183] 可以理解的，第二目标向量为最终查询向量对应的目标向量。

[0184] 可以理解的，在第一目标向量中有相同的特征向量时，将其中一个舍去。

[0185] 本申请提供的向量搜索的方式，将查询向量分发到各个节点中，可以获取查询向量距离各节点内的聚类集合的代表点即中心点的距离，并可以获取查询向量与各节点的代表点之间的距离，并获取查询向量在各节点内对应的目标向量。然后可以将查询向量在各节点内对应的目标向量汇总进行对比，以实现搜索到与查询向量距离更近的，或更为精确的目标向量，能够有效提高搜索精度。

[0186] 可以理解，本申请实施例中将索引构建方法应用于上述图片数据库中的索引构建只是举例说明，本申请实施例中提供的索引构建方法可以应用于各种视频、语音、蛋白质分子结构等数据库中，即本申请实施例中提供的的索引构建方法可以广泛应用于图像、视频、语音、蛋白质分子结构检索等领域中。

[0187] 可以理解，本申请实施例中提及的检索系统可以包括至少一个数据库，例如，可以包括上述图片数据库，还可以包括视频数据库、文档数据库等。其中，图片数据库包括有多个原始图片数据对应的特征向量，视频数据库包括有多个原始视频数据对应的特征向量，文档数据库包括有多个原始文档数据对应的特征向量。可实施的，每个数据库均可以采用上述索引构建方法进行索引的构建。

[0188] 下面对本申请实施例提供的索引构建方法进行详细叙述。可以理解，本申请实施例提供的索引构建方法可以由检索系统执行，也可以由其他电子设备执行，即其他电子设备将对各个数据库进行索引的构建，然后将构建好的索引的各个数据库部署至检索系统

中。

[0189] 下面以索引构建方法由检索系统执行为例对本申请实施例提供的索引构建方法进行详细叙述，其中，以系统中有3个节点，节点次序依次为第一节点、第二节点和第三节点为例，图8示出了本申请实施例一种索引构建方法的流程示意图。

[0190] 如图8所示，本申请实施例中的索引构建方法可以包括：

[0191] 801：获取各个节点中待创建索引的数据库中各数据所对应的各特征向量。

[0192] 可以理解，所述数据库可以为包括有各种图片、视频、语音、蛋白质分子结构等数据结构中的一种或多种的数据库。由于各种数据均可以被转化为高维度的特征向量，因此，检索系统可以首先将数据库中的各原始数据转化为对应的特征向量。可以理解，数据库中的各原始数据也可以仍然保留并存储于数据库中。

[0193] 例如，检索系统可以首先获取如图1a所示的数据库S中通过相同的图片数据库中的图片数据转化为的特征向量。

[0194] 802：将获取的各特征向量冗余(重复)放置在预设数量的节点中。

[0195] 可以理解的，预设数量的节点可以为大于等于两个的节点。

[0196] 例如，将获取的如图1a所示的数据库S中的图片数据转化为的特征向量冗余(重复)放置在3个节点中，例如放在如图2所示的第一节点、第二节点以及第三节点中。可以理解的，这3个节点中的特征向量相同。

[0197] 803：预设第一节点内的代表点，并基于预设的代表点对第一节点内的特征向量进行聚类处理，获取对应数量的聚类集合，其中每个聚类集合具有对应的代表点。

[0198] 对第一节点预先设置对应数量个代表点，并使用预先设置的代表点对该节点内的特征向量进行聚类处理，获取多个聚类集合以及每个聚类集合对应的代表点。

[0199] 可以理解，在一些实施例中，根据任意节点的预设代表点对节点内特征向量进行聚类后获得对应的聚类集合后，可以重新计算各聚类集合的代表点。例如，将各聚类集合中各特征向量的算术平均值作为该聚类集合的代表点。可以理解，重新计算的各聚类集合的代表点与初始预设的代表点一般为不一致的。但本申请实施例为了方便描述，本申请的实施例中以预设代表点近似替代聚类后的代表点为例对本申请实施例中的技术方案进行介绍。

[0200] 在一些实现方式中，根据预先设置的代表点对第一节点中的特征向量进行聚类处理，将各特征向量分别分配至对应的聚类集合中，每个聚类集合具有对应的代表点。可以理解，对各特征向量进行聚类处理的方式可以为k-means聚类方式、Agglomerative算法等。

[0201] 例如，以使用k-means聚类方式、第一预设数量为4为例对第一节点的聚类进行说明。如图2所示，在第一节点中，为图片数据库S预先设置4个预设代表点，可以理解，第一节点的预设代表点可以是随机产生的，可实施的，可以是多个距离尽可能远的点。然后根据第一节点中每个特征向量与各聚类集合的预设代表点的距离将各特征向量分别分配至距离最近的预设代表点所对应的聚类集合，重新确定各聚类集合对应的代表点，获取如图3a所示的多个聚类集合S1、S2、S3、S4以及每个聚类集合对应的代表点C1、C2、C3、C4。可以理解的，可以将上述第一节点中的各聚类集合对应的预设代表点作为聚类集合所对应的代表点，也可以将该聚类集合中所有特征向量的算术平均向量确定为该聚类集合的代表点，或则以其他方式确定该聚类集合中的代表点。可以理解，本申请实施例中所提及的距离可以

为欧式距离,也可以为内积距离等其他距离。

[0202] 可以理解,上述在第一节点内对所述各特征向量进行聚类处理的方式也可以为其他聚类方式。可以理解的,上述聚类处理的方式还可以为Agglomerative等聚类方式。

[0203] 804:从第一节点的各聚类集合中,获取第一预设数量的边缘点,作为第二节点的各聚类集合的预设代表点,基于第二节点的各聚类集合的预设代表点对第二节点内的特征向量进行聚类处理,获取对应的第一预设数量个聚类集合,每个聚类集合具有对应的代表点。

[0204] 可以理解的,第二节点中经过聚类处理之后获得的各聚类集合对应的代表点可以与预设的代表点,即第一节点的各边缘向量相同或者不同。其中,经过聚类处理之后获得的各聚类集合对应的代表点与预设的代表点不同,可以为各聚类集合中所有特征向量对应的的平均向量,或则以其他方式确定该聚类集合中的代表点。

[0205] 可以理解,在第二节点及其他节点对所述各特征向量进行聚类处理的方式均可以采用上述提及的k-means聚类方式、Agglomerative算法等。

[0206] 可以理解,在确定第二节点的聚类集合的总数后,即第一预设数量确定后,可以根据以下方法确定第一节点各聚类集合内需要确定的边缘点的第二预设数量:

[0207] 第一种可实施的方案中,确定第一节点中各聚类集合的边缘点的第二预设数量的方法为:根据第一预设数量,按照第一节点中各聚类集合的半径,根据第一预设数量以及各个聚类集合的半径的比例确定每个聚类集合的边缘点的第二预设数量;可以理解,在一些实施例中,当计算出来的任一聚类集合的边缘点的预设数量不为整数时,可以根据取整规则获取最终的预设数量。例如,取整规则可以为采取将当前数量对应的数值的整数位加1,舍去小数部分所获得的数值,作为最终的的预设数量所对应的数值。

[0208] 例如,第一预设数量为4,即边缘点预设总数量为4,如图3a所示的第一节点对应的聚类集合中,假设聚类集合S1半径大小为16、聚类集合S2半径大小为14、聚类集合S3半径大小为10以及聚类集合S4半径大小为17,则聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的的数量比例为16:14:10:17;根据边缘点的预设总数量和上述边缘点的数量比例为16:14:10:17可以确定出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的数量为1.1,0.9,0.7和1.7。根据上述取整规则,可以得出聚类集合S1、聚类集合S2、聚类集合S3、聚类集合S4中应该确定的边缘点的第二预设数量分别为2,1,1和2。

[0209] 可以理解,本申请实施例中,通过将各节点中预设数量的边缘点作为下一节点的预设代表点的方式可以使得各节点的聚类尽可能不同,如此,可以有效提高搜索精度。

[0210] 其次,上述根据第一预设数量以及各个聚类集合的半径的比例确定每个聚类集合的边缘点的预设数量的方式可以使得边缘点的分布更加均匀,避免出现有些聚类集合半径很小,但还确定了数量较多的边缘点的情况的发生。

[0211] 第二种可实施的方案中,确定第一节点中各聚类集合的边缘点的第二预设数量的方法为:根据第一预设数量,按照第一节点中各聚类集合内半径从大到小排序,从每个聚类集合中选取设定个数的边缘点,直到选完预设第一数量个。

[0212] 例如,第一预设数量为4,即边缘点预设总数量为4,如图3a所示的第一节点对应的聚类集合中,假设聚类集合S1半径大小为16、聚类集合S2半径大小为14、聚类集合S3半径大

小为10以及聚类集合S4半径大小为17，则按照各个聚类集合半径大小排列得到的顺序为聚类集合S4-聚类集合S1-聚类集合S2-聚类集合S3；当每个聚类集合中选取的边缘点的设定个数为2个，按照聚类集合S4-聚类集合S1-聚类集合S2-聚类集合S3的先后顺序，聚类集合S4和聚类集合S1中需要确定的边缘点的各第二预设数量均为两个。因为此时已达到边缘点的预设总数量。后续聚类集合S2和聚类集合S3将不再选取边缘点，即排在聚类集合S1后面的聚类集合S2和聚类集合S3的个数为0个。

[0213] 可以理解，上述根据第一预设数量，按照各个聚类集合内特征向量的数量或半径从大到小排序，从每个聚类集合中选取设定个数的边缘点的方式可以使得在较小聚类集合中确定数量较少的边缘点，而在其他较大的聚类集合确定较多的边缘点，可以使得确定的边缘点能够充分代表对应的聚类集合，有效提高搜索精度。

[0214] 可以理解，确定其他节点内的各聚类集合所需要确定的边缘点的第二预设数量的方式均可以采用上述方式。

[0215] 在基于上述方式确定各节点，例如第一节点内的各聚类集合所需要确定的边缘点的第二预设数量后，则需要在对应聚类集合内确定对应的第二预设数量的边缘点。下面以在第一节点内的第一聚类集合中确定第二预设数量的边缘点的方式为例，对本申请实施例中提供的在任意聚类集合中确定第二预设数量的边缘点的方式进行介绍：

[0216] 第一种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：获取第一聚类集合的代表点与该第一聚类集合内的所有特征向量的第一距离；按照第一距离将各特征向量由远到近排序，得到第一聚类集合对应的第一序列；确定第一序列中前第二预设数量个特征向量为该聚类集合内的边缘点。

[0217] 可以理解，本申请实施例中，可以根据第一聚类集合中各特征向量与代表点之间的距离确定边缘点，便于能够将距离代表点较远的特征向量作为边缘点，使得边缘点的选取更加精确。

[0218] 且本申请实施例中，根据第一聚类集合中各特征向量与代表点之间的距离，即根据各特征向量与代表点的远近情况，将各特征向量从远到近排序，可以使得根据设定数量选取对应的边缘点时更加方便。例如，边缘点的设定数量为五个，则可以直接将排序即第一序列中的前五个特征向量作为边缘点。

[0219] 第二种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：获取第一聚类集合的代表点与该第一聚类集合内所有特征向量的第一距离；将所述第一距离大于设定距离的特征向量作为第一序列中的特征向量，然后将第一序列中的第二预设数量的特征向量作为边缘点，或者将第一序列中的各特征向量按照第一距离进行由远至近的排序，将排序中的前第二预设数量的特征向量作为边缘点。

[0220] 可以理解，本申请实施例中，可以根据第一聚类集合中各特征向量与代表点之间的距离大于设定距离的特征向量获得第一序列，即第一序列中的特征向量是距离代表点较远的且大于设定距离的向量，且可以使得选取相邻聚类集合的方式更加规范且简单。例如，可以只需在选取边缘点的算法中设置相应的距离阈值参数即可实现边缘点的获取。

[0221] 第三种可实施的方案中，从第一聚类集合中确定第二预设数量的边缘点的方式可以为：确定除第一聚类集合之外的其他聚类集合中，与第一聚类集合邻近的至少一个相邻聚类集合；获取第一聚类集合中每个特征向量分别与每个相邻聚类集合的代表点的距离；

根据该距离，对各特征向量由近至远排序，获取每个相邻聚类集合对应的第一序列；将第一序列中的前设定数量个特征向量作为边缘点。

[0222] 其中，获取第一聚类集合的邻近的至少一个相邻聚类集合的方式可以为：将上述多个聚类集合中除第一聚类集合之外的各聚类集合，按照各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行由近至远的排序，获取该第一聚类集合所对应的第二序列；将所述第二序列中的前第二预设数量个聚类集合作为与第一聚类集合邻近的相邻聚类集合。

[0223] 下面以前述图3a中的第一聚类集合S1为例，介绍获取第一聚类集合S1中的边缘点的方式：

[0224] 首先可以获取第一聚类集合S1的邻近的至少一个相邻聚类集合。其中获取第一聚类集合S1的邻近的至少一个相邻聚类集合的方式可以为：获取与除聚类集合S1之外的其他三个聚类集合的代表点与第一聚类集合S1的代表点C1之间的距离。例如，如图4所示，可以获取到聚类集合S2的代表点C2与第一聚类集合S1的代表点C1与的距离为d5，聚类集合S3的代表点C3与第一聚类集合S1的代表点C1与的距离为d6，聚类集合S4的代表点C4与第一聚类集合S1的代表点C1与的距离为d7，然后根据上述三个聚类集合各自的代表点与第一聚类集合S1的代表点C1之间的距离由近到远，对上述三个聚类集合进行排序，获取代表点C1对应的第二序列；假设d7＜d5＜d6时，则代表点C1对应的第二序列为聚类集合S4-聚类集合S2-聚类集合S3；此时，将代表点C1与所述第二序列中的前第二预设数量个聚类集合建立索引关联。假设第二预设数量为2个，则可以将聚类集合S2、聚类集合S4作为第一聚类集合S1的邻近的相邻聚类集合。

[0225] 然后，获取第一聚类集合S1中每个特征向量与相邻聚类集合S2的代表点C2的距离，以及第一聚类集合S1中每个特征向量与相邻聚类集合S4的代表点C2的距离。如图5a中所示，对于特征向量X1，获取第一聚类集合S1中特征向量X1与相邻聚类集合S2的代表点C2的距离d2_1，对于其他特征向量与代表点C2的距离在此不再赘述；根据各特征向量与代表点C2的距离，对各特征向量由近至远排序，假设根据距离，将第一聚类集合S1中15个特征向量由远至近排序得到的第一序列为X1-X3-X5-.....-X15。假设预设的根据每个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个，则可以将相邻聚类集合S2对应的第一序列中的第一个特征向量，即特征向量X1作为边缘点。同时，获取第一聚类集合S1中每个特征向量与相邻聚类集合S4的代表点C4的距离，如图3a中所示，对于特征向量X1，获取第一聚类集合S1中特征向量X1与相邻聚类集合S4的代表点C4的距离d4_1，对于其他特征向量与代表点C4的距离在此不再赘述；根据各特征向量与代表点C4的距离，对各特征向量由近至远排序，假设根据距离，将第一聚类集合S1中15个特征向量由近至远排序得到的第一序列为X7-X8-X1-.....-X14。假设预设的根据每个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个，则可以将第四聚类集合S4对应的第一序列中的第一个特征向量，即特征向量X7作为边缘点。

[0226] 可以理解的，当第二距离越小就说明该特征向量越接近相邻聚类集合，也就是越接近该第一聚类集合的边缘。因此，上述通过第一聚类结合中各特征向量与各相邻聚类集合的距离获取边缘点的方式，便于获取第一聚类集合在与靠近各相邻聚类集合方向上的边缘点，将第一聚类集合在与靠近各相邻聚类集合方向上的边缘点作为边缘点，便于在检索

时，能够根据边缘点和代表点与查询向量的距离确定出与查询向量最为接近的代表点，并对该代表点对应的聚类集合进行搜索，对于与各聚类集合的代表点的距离较远，而与边缘点即边缘点的距离较近的查询向量，在查询的过程中能够增大目标向量被查找到的概率，有效提升搜索精度。

[0227] 可以理解，上述通过按照各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行排序来获取第一聚类集合的邻近聚类集合，即相邻聚类集合的方式只是示例性说明。可实施的，本申请实施例获取第一聚类集合的邻近聚类集合即相邻聚类集合的方式也可以为其他方式。

[0228] 其中，获取相邻聚类集合的方式还可以为：将上述多个聚类集合中除第一聚类集合之外的各聚类集合的代表点与第一聚类集合的代表点之间的距离小于设定距离的聚类集合组成第二序列中的聚类集合；然后将第二序列中的任意第二预设数量的聚类集合作为相邻聚类集合，或者将第二序列中的各聚类集合的代表点与该第一聚类集合的代表点之间的距离进行由近至远的排序，将排序中的前第二预设数量的聚类集合作为相邻聚类集合。

[0229] 例如，如图4所示，确定与聚类集合S1邻近的相邻聚类集合的方式可以为：首先获取除聚类集合S1之外的其他聚类集合的代表点与聚类集合S1的代表点之间的距离。例如，聚类集合S2的代表点C2与第一聚类集合S1的代表点C1与的距离为d5，聚类集合S3的代表点C3与第一聚类集合S1的代表点C1与的距离为d6，聚类集合S4的代表点C4与第一聚类集合S1的代表点C1与的距离为d7。然后将d5、d6以及d7与设定距离D进行比较，假设d5、d7小于设定距离D，则聚类集合S2、聚类集合S4组成第二序列中的聚类集合；然后将第二序列中的聚类集合S2、聚类集合S4中的代表点与该聚类集合S1的代表点之间的距离d5、d7进行由近至远的排序，假设d7小于d5，则排序为聚类集合S4-聚类集合S2。若相邻聚类集合的第二预设数量为一个，则将聚类集合S4作为相邻聚类集合。

[0230] 再例如，获取相邻聚类集合的方式还可以为：估算除第一聚类集合的代表点本身所在的第一聚类集合之外的各聚类集合的近似半径距离，获取各聚类集合的代表点与第一聚类集合的代表点的之间的距离与近似半径距离的差值，根据差值，将各聚类集合进行由近至远排序获取第二序列，将第二序列中的前第二预设数量个聚类集合作为相邻聚类集合。

[0231] 可以理解，在一些实施例中，上述估算除目标特征向量本身所在的第一聚类集合之外的各聚类集合的近似半径距离的方式可以为：获取各聚类集合中所有特征向量与各聚类集合中的代表点之间的距离，将所有特征向量与代表点之间的最大距离作为各聚类集合的近似半径距离。

[0232] 例如，如图5b中所示，确定与聚类集合S1邻近的相邻聚类集合的方式可以为：首先，确定除聚类集合S1之外的各聚类集合S2、S3以及S4，将聚类集合S2中各特征向量中与聚类集合S2的代表点C2的最远距离作为聚类集合S2的近似半径距离，假设最远距离为特征向量Y3与聚类集合S2的代表点C2之间的距离r1，则将r1作为聚类集合S2的近似半径距离；将聚类集合S3中各特征向量中与第一中心点C3的最远距离作为聚类集合S2的近似半径距离，假设最远距离为特征向量Z1与聚类集合S3的代表点C3之间的距离r2，则将r2作为聚类集合S3的近似半径距离；将聚类集合S4中各特征向量中与第一中心点C4的最远距离作为聚类集合S4的近似半径距离，假设最远距离为特征向量W5与聚类集合S4的代表点C4之间的距离

r3,则将r3作为聚类集合S4的近似半径距离r3。

[0233] 然后,确定聚类集合S2的代表点C2与聚类集合S1的代表点C1与的距离d5与聚类集合S2的半径距离r1的差值为d5-r1,聚类集合S3的代表点C3与聚类集合S1的代表点C1与的距离d6与聚类集合S3的半径距离r2的差值为d6-r2,聚类集合S4的代表点C4与聚类集合S1的代表点C1与的距离d7与聚类集合S4的半径距离r3的差值为d7-r3,根据差值,将聚类集合S2、S3以及S4进行由近至远排序获取第二序列,假设d7-r3<d5-r1<d6-r2,则第二序列为聚类集合S4-聚类集合S2-聚类集合S3,将第二序列中的前第二预设数量个聚类集合作为相邻聚类集合,假设将第二序列中的前1个数量的聚类集合作为相邻聚类集合,则相邻聚类集合为聚类集合S4。

[0234] 在一些实施例中,估算各聚类集合的近似半径的方式也可以为其他任意可实施的方式,例如通过神经网络模型或者相关算法估算等。

[0235] 可以理解,任一聚类集合的代表点与第一聚类集合的代表点之间的距离与近似半径的差值可以近似代表该聚类集合中的边缘特征向量与第一聚类集合的代表点之间。当边缘特征向量距离第一聚类集合的代表点之间的距离较近,则可能存在该边缘特征向量所在聚类集合为第一聚类集合的邻近集合。

[0236] 第四种可实施的方案中,确定符合预设条件的至少一个边缘点的方式可以为:确定除第一聚类集合之外的其他聚类集合中,与所述第一聚类集合邻近的至少一个相邻聚类集合;获取第一聚类集合中每个特征向量分别与所述第一聚类集合的代表点之间的第一距离以及与所述每个相邻聚类集合的代表点之间的第二距离,并根据第一距离和第二距离的距离之和;对各特征向量,按照距离之和由远至近排序,获取每个相邻聚类集合对应的所述第一序列;将第一序列中的前设定数量个特征向量作为边缘点。可以理解,第四种方案中获取第一聚类集合邻近的至少一个相邻聚类集合的方式与第三种方案中阐述的获取相邻聚类集合的方式相同,此处不在赘述。

[0237] 可以理解的,根据第一距离和第二距离的距离之和从大到小进行排序得到第一序列,得到的第一序列中距离之和较大的特征向量,即为距离第一聚类集合的代表点较远,且距离相邻聚类集合的代表点也较远的特征向量,即该特征向量可能位于代表点的代表点指向边缘点的相反方向上的边缘点。当获取了第一聚类集合的各方向上的相邻聚类集合对应的第一序列,则可以获取第一聚类集合的各方向上的边缘点。如此,将第一聚类集合的各方向上的边缘点作为边缘点,可以有效提高搜索精确度。

[0238] 下面以获取图3a中第一聚类集合S1的至少一个边缘点为例,说明本申请实施例中第四种实施方案中获取边缘点的方式:

[0239] 如图5c中所示,假设第一聚类集合S1邻近的相邻聚类集合为前述图4中所述的聚类集合S2和聚类集合S4。根据上述确定的第一聚类集合S1的邻近的相邻聚类集合S2,获取第一聚类集合S1中每个特征向量分别与该第一聚类集合S1的代表点C1之间的第一距离,以及与相邻聚类集合S2的代表点C2的第二距离的距离之和。

[0240] 例如,如图5c中所示,第一聚类集合S1中的特征向量X1与代表点C1之间的第一距离d1_1以及与代表点C2的第二距离d2_1的距离之和为d_s2,其他14个特征向量获得距离之和的方式相同,在此不再赘述;假设根据距离之和将第一聚类集合S1中15个特征向量由远至近排序得到的聚类集合S2对应的第一序列为X2-X3-X1-.....-X14,假设预设的根据每

个邻近集合的获取的对应聚类集合的边缘点的第二预设数量为一个,则可以将相邻聚类集合S2对应的第一序列中的第一个特征向量,即特征向量X2作为边缘点。

[0241] 同时,根据上述确定的第一聚类集合S1的邻近的相邻聚类集合S4,获取第一聚类集合S1中每个特征向量分别与该第一聚类集合S1的代表点C1之间的第一距离以及与相邻聚类集合S4的代表点C4的第二距离的距离之和,例如,图4中,第一聚类集合S1中的特征向量X1与代表点C1的第一距离d1_1以及代表点C4的第二距离d4_1的距离之和d_s4,其他14个特征向量获得距离之和的方式相同,在此不再赘述;假设根据距离之和将第一聚类集合S1中15个特征向量由远至近排序得到的第一序列为X4-X1-X3-.....-X15,假设预设的根据每个邻近集合的获取的对应聚类集合的预设边缘点的数量为一个,则可以将第四聚类集合S4对应的第一序列中的第一个特征向量,即特征向量X4作为边缘点。

[0242] 第五种可实施的方案中,从第一聚类集合中确定第二预设数量的边缘点的方式可以为:确定除第一聚类集合之外的其他聚类集合中,与所述第一聚类集合邻近的至少一个相邻聚类集合;获取第一聚类集合中每个特征向量分别与所述第一聚类集合的代表点之间的第一距离以及与所述每个相邻聚类集合的代表点之间的第二距离。并根据第二距离和第一距离的差值与第一距离的比值,对第一聚类集合中各特征向量,按照比值从小到大排序,获取每个相邻聚类集合对应的第一序列;将第一序列中的前第二预设数量个特征向量作为边缘点。可以理解,第五种方案中获取该聚类集合邻近的至少一个相邻聚类集合的方式与第三种方案中阐述的获取相邻聚类集合的方式相同,此处不在赘述。

[0243] 可以理解的,当特征向量与相邻聚类集合的代表点之间的第二距离越小,,特征向量与第一聚类集合的代表点之间的第一距离越大,则证明特征向量距离第一聚类集合的代表点越远,即第二距离和第一距离的差值与第一距离的比值越小。因此,按照第二距离和第一距离的差值与第一距离的比值由小到大对特征向量进行排序,可以获得第一聚类集合的更为精确的边缘点。

[0244] 例如,下面以获取图3a中聚类集合S1的第二预设数量的边缘点为例,说明本申请实施例中第五种实施方案中获取边缘点的方式:

[0245] 假设聚类集合S1邻近的相邻聚类集合为前述图4中所述的聚类集合S2和聚类集合S4。根据上述确定的聚类集合S1的邻近的相邻聚类集合S2,获取聚类集合S1中每个特征向量分别与该聚类集合S1的代表点C1之间的第一距离,以及与相邻聚类集合S2的代表点C2的第二距离的距离之和。

[0246] 例如,如图5d中所示,聚类集合S1中的特征向量X1与代表点C2之间的第二距离d2_1以及与代表点C1的第一距离d1_1的距离差值与第一距离d1_1的比值为d_b2,其他14个特征向量获得距离之比的方式相同,在此不再赘述;假设根据比值由小至大将聚类集合S1中15个特征向量排序得到的聚类集合S2对应的第一序列为X8-X3-X1-.....-X14,假设预设的根据每个邻近集合的获取的对应聚类集合的相邻代表点的预设数量为一个,则可以将相邻聚类集合S2对应的第一序列中的第一个特征向量,即特征向量X8作为边缘点。

[0247] 同时,根据上述确定的聚类集合S1的邻近的相邻聚类集合S4,获取聚类集合S1中的特征向量X1与代表点C4之间的第二距离d4_1以及与代表点C1的第一距离d1_1的距离差值与第一距离d1_1的比值为d_b4,其他14个特征向量获得比值的方式相同,例如,图5d中,聚类集合S1中的特征向量X1与代表点C4的第二距离d4_1以及代表点C1的第一距离d1_1的

距离差值与第一距离d1_1的比值为d_b4，其他15个特征向量获得比值的方式相同，在此不再赘述；假设根据比值由小到大将聚类集合S1中15个特征向量排序得到的第一序列为X9-X1-X3-.....-X15，假设预设的根据每个邻近集合的获取的对应聚类集合的预设第二代表点的数量为一个，则可以将聚类集合S4对应的第一序列中的第一个特征向量，即特征向量X9作为边缘点。

[0248] 805：判断是否所有节点均已进行聚类处理。

[0249] 在一些实现方式中，若预设节点的数量为两个，即所有节点均已进行聚类处理，则转至步骤806，确定各目标向量在各节点内所在的聚类集合；将各目标向量与各目标向量在各节点内所在的聚类集合对应的代表点建立索引关联。

[0250] 若预设节点的数量不为两个，即还存在未进行聚类处理的节点，则转至步骤807，从第二节点的各聚类集合中，获取第三预设数量的边缘点，作为第三节点的各聚类集合的预设代表点，基于第三节点的各聚类集合的预设代表点对第三节点内的特征向量进行聚类处理，获取第三预设数量个聚类集合，并建立第三节点对应的索引关联，其中，第三节点中每个聚类集合具有对应的代表点。

[0251] 806：建立各特征向量与各节点对应的代表点之间的索引关联。

[0252] 在一些实现方式中，可以确定目标向量在所有节点，例如第一节点以及第二节点所在的聚类集合；将目标向量与目标向量在所有节点，例如第一节点以及第二节点所在的聚类集合对应的代表点建立索引关联。

[0253] 例如，对于如图3a所示的第一节点，可以将该目标向量与该目标向量在第一节点中所在的聚类集合对应的代表点在子索引结构中建立索引关联。如图9所示，第一节点对应的子索引结构包括代表点项和倒排文件项。代表点项包括聚类集合S1具有对应的代表点C1，聚类集合S2具有对应的代表点C2，聚类集合S3具有对应的代表点C3，和聚类集合S4具有对应的代表点C4。

[0254] 聚类集合S1的对应的代表点C1具有对应的倒排文件D1，该倒排文件D1中包括对应的代表点C1所对应的聚类集合S1中的各个特征向量。

[0255] 聚类集合S2的对应的代表点C2具有对应的倒排文件D2，该倒排文件D2中包括对应的代表点C2所对应的聚类集合S2中的各个特征向量。

[0256] 聚类集合S3的对应的代表点C3具有对应的倒排文件D3，该倒排文件D3中包括对应的代表点C3所对应的聚类集合S3中的各个特征向量。

[0257] 聚类集合S4的对应的代表点C4具有对应的倒排文件D4，该倒排文件D4中包括对应的代表点C4所对应的聚类集合S4中的各个特征向量。

[0258] 图9也示出了如图3b所示的第二节点对应的子索引结构，该子索引结构与第一节点的子索引结构类似，在此不做赘述。

[0259] 807：从第二节点的各聚类集合中，获取第三预设数量的边缘点，作为第三节点的各聚类集合的预设代表点，基于第三节点的各聚类集合的预设代表点对第三节点内的特征向量进行聚类处理，获取第三预设数量个聚类集合，其中，第三节点中每个聚类集合具有对应的代表点。

[0260] 可以理解的，从第二节点的各聚类集合中，获取第三预设数量的边缘点，作为第三节点的各聚类集合的预设代表点，基于第三节点的各聚类集合的预设代表点对第二节点内

的特征向量进行聚类处理后得到的第三节点的各聚类集合的方法，与上述步骤804中确定第二节点的各聚类集合的方法相同，具体方式在此不做赘述。

[0261] 例如，根据上述步骤804中确定第二节点的各聚类集合的方法得到如图3c所示的第三节点中的各聚类结合，聚类集合S1”、聚类集合S2”以及聚类集合S3”。

[0262] 图10示出了本申请实施例中一种搜索方法的流程示意图，该搜索方法可以用于各种包括上述索引结构的数据库，该搜索方法可以由包括上述数据库的检索系统执行。如图10所示，本申请实施例中的搜索方法可以包括：

[0263] 1001：获取查询向量。

[0264] 可以理解，用户在进行信息检索时，可以在检索系统的搜索窗口或搜索框中输入对应的查询数据，检索系统在获取到该查询数据后，可以将该查询数据转化为对应的查询向量。可以理解，用户输入的查询数据可以为任意格式的查询数据。

[0265] 例如，用户输入的查询数据可以为图片格式，在检索系统在获取到该图片后，可以将该图片转化为对应的查询向量。

[0266] 例如，用户输入的查询数据可以为文本格式，在检索系统在获取到该文本后，可以将该文本转化为对应的查询向量。

[0267] 例如，用户输入的查询数据可以为视频格式，在检索系统在获取到该视频后，可以将该视频转化为对应的查询向量。

[0268] 1002：获取各节点中各聚类集合的第一代表点与查询向量的第三距离。

[0269] 例如，基于上述构建完成的如图3d所示的索引关联，检索系统可以先获取查询向量A，将查询向量A分发到第一节点、第二节点以及第三节点中。然后分别确定出第一节点中各第一聚类集合的代表点与查询向量A之间的第三距离、第二节点中各第二聚类集合的代表点与查询向量A之间的第三距离以及第三节点中各第三聚类集合的代表点与查询向量A之间的第三距离。

[0270] 例如图6a示出了获取查询向量A与第一节点中聚类集合S1的代表点C1、聚类集合S2的代表点C2、聚类集合S3的代表点C3以及聚类集合S4的代表点C4的第三距离分别为da1、da2、da3以及da4。例如图6b示出了获取查询向量A与第二节点中聚类集合S1’的代表点C1’、聚类集合S2’的代表点C2’、聚类集合S3’的代表点C3’以及聚类集合S4’的代表点C4’的第三距离分别为dm1、dm2、dm3以及dm4。例如图6c示出了获取查询向量A与第三节点中聚类集合S1”的代表点C”、聚类集合S2”的代表点C2”、聚类集合S3”的代表点C3”的第三距离分别为dk1、dk2以及dk3。

[0271] 1003：根据第三距离，确定出各节点内的目标代表点，并确定目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离。

[0272] 在一些实现方式中，可以将各节点中的各代表点中，与查询向量之间的第三距离在设定距离内的代表点作为目标代表点。

[0273] 在一些实现方式中，还可以将各节点中的各代表点按照与查询向量之间的第三距离进行排序，在排序的前预设数量个代表点作为目标代表点，并确定获得的目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离。

[0274] 例如，根据如图3a-3c所示的3个节点中查询向量A与每个聚类集合中各代表点之间的第三距离，确定出各节点中与查询向量距离目标代表点，在各节点中选择距离最近的

一个代表点作为各节点的目标代表点，假设第一节点中与查询向量A距离最近的代表点为C1，第二节点中与查询向量A距离最近的代表点为C1'，第三节点中与查询向量A距离最近的代表点为C2'，则将第一节点中与查询向量A距离最近的代表点C1，第二节点中与查询向量A距离最近的代表点C1'，第三节点中与查询向量A距离最近的代表点C2'为各节点对应的目标代表点。根据如图3d所示的索引关联，获取第一节点中代表点为C1所在聚类集合S1、第二节点中代表点为C1'所在聚类集合S1'、第三节点中代表点为C2"所在聚类集合S2"中各特征向量与查询向量的第四距离。

[0275] 1004:根据第四距离，确定出查询向量在各节点中对应的第一目标向量。

[0276] 在一些实施例中，可以将各特征向量，按照第四距离由近至远进行排序，将排序的前预设数量个特征向量作为第一目标向量。

[0277] 在一些实施例中，可以将与查询向量A之间的第四距离小于设定值的特征向量作为第一目标向量。

[0278] 例如，将上述步骤中获取的第一节点中聚类集合S1、第二节点中聚类集合S1'以及第三节点中聚类集合S2"中与查询向量A之间的距离在设定范围内的特征向量作为查询向量A在各节点中对应的目标向量，

[0279] 如图7a所示，假设第一节点中聚类集合S1中与查询向量A在设定范围内的特征向量为特征向量X1、特征向量X2且对应的第四距离分别为dn1、dn2，如图7b所示，第二节点中聚类集合S1'中与查询向量A在设定范围内的特征向量为特征向量X1'、特征向量X2'、特征向量X3'且对应的第四距离分别为dt1、dt2、dt3，如图7c所示，第三节点中聚类集合S2"中与查询向量A在设定范围内的特征向量为特征向量Y1'、特征向量Y2'且对应的第四距离分别为dw1、dw2。则第一节点对应的特征向量X1、特征向量X2，第二节点对应的特征向量X1'、特征向量X2'、特征向量X3'以及第三节点对应的特征向量Y1'、特征向量Y2'中为查询向量对应的的第一目标向量。

[0280] 1005:按照第四距离，从查询向量在各节点中对应的第一目标向量中，确定查询向量对应的第二目标向量。

[0281] 可以理解的，第二目标向量为最终查询向量对应的目标向量。

[0282] 在一些实现方式中，可以按照所述第四距离由近至远，将各节点中对应的第一目标向量进行排序，获得排序中的前预设数量个特征向量作为第二目标向量。

[0283] 例如，根据获取的第一目标向量与查询向量A对应的第四距离，可以对各第一目标向量进行排序，获得排序中的前预设数量个特征向量作为第二目标向量

[0284] 假设在图7a中确定出的第一目标向量为第一节点对应的特征向量X1、特征向量X2，第二节点对应的特征向量X1'、特征向量X2'、特征向量X3'以及第三节点对应的特征向量Y1'、特征向量Y2'。且上述特征向量与查询向量之间的距离分别为dn1、dn2、dt1、dt2、dt3、dw1、dw2，假设dw1<dn1<dn2<dt1<dt2<dt3<dw2，则第一目标向量对应的排序为特征向量Y1'-特征向量X1-特征向量X2、特征向量X1'-特征向量X2'-特征向量X3'-特征向量Y2'，假设第二目标向量的预设数量为两个，确定第三节点中的Y1"、以及第一节点中的X1为第二目标向量。

[0285] 可以理解，本申请实施例中，当检索系统获取到目标向量后，可以将目标向量对应的原始数据输出至客户端。

[0286] 例如，当数据库为图片数据库，则检索系统可以将目标向量对应的原始图片数据输出至客户端。

[0287] 可以理解，上述搜索方法可以用于对检索系统中的3个节点进行搜索，也可以用于对检索系统中的多个节点进行搜索。

[0288] 当检索系统中包括节点时，检索系统可以基于查询向量对该多个节点中执行上述图10中所示的搜索方法，并获取对应的搜索结果，然后将所有的搜索结果输出。

[0289] 可以理解，检索系统获取的输入数据以及检索系统确定出的输出数据的格式可以相同，也可以不同。输出数据格式与被搜索数据库所包含的数据格式相关，例如，当被搜索的数据库中的数据格式包括文本格式、图片格式或视频格式。则输出数据格式存在文本格式、图片格式或视频格式的可能性。

[0290] 图11示出了本申请实施例一种索引构建装置的示意图，如图所示，索引构建装置包括：

[0291] 第一确定单元，用于确定目标向量以及目标向量在当前节点内所在的第一聚类集合，其中第一聚类集合具有第一代表点；

[0292] 第二确定单元，用于确定所述目标向量在下一节点内所在的第二聚类集合，所述第二聚类集合具有对应的代表点，所述第二聚类集合对应的预设代表点为所述第一节点内其中一个聚类集合的边缘点；其中，所述第一节点内的特征向量与所述第二节点内的特征向量相同；

[0293] 关联单元，分别在各所述节点中，建立所述目标向量与所述目标向量所在的第一聚类集合之间的索引关联。。

[0294] 图12示出了本申请实施例一种搜索装置的示意图，如图12所示，搜索装置包括：

[0295] 第一获取单元，用于获取查询向量；

[0296] 第二获取单元，用于获取各节点中各聚类集合的代表点与查询向量的第三距离；

[0297] 第一确定单元，用于根据所述第三距离，确定出各节点内的目标代表点，并确定所述目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第二距离；

[0298] 第二确定单元，用于根据所述第四距离，确定出所述查询向量在所述各节点中对应的第一目标向量；

[0299] 第三确定单元，用于按照所述第四距离，从所述查询向量在所述各节点中对应的第一目标向量中，确定所述查询向量对应的第二目标向量。

[0300] 如图13所示，本申请实施例中还包括一种检索系统，包括上述索引构建装置，和至少一个数据库和搜索装置。

[0301] 图14示出了本申请实施例一种电子设备的框图。在一个实施例中，电子设备1400可以包括一个或多个处理器1404，与处理器1404中的至少一个连接的系统控制逻辑1408，与系统控制逻辑1408连接的系统内存1412，与系统控制逻辑1408连接的非易失性存储器(NVM)1416，以及与系统控制逻辑1408连接的网络接口1420。

[0302] 在一些实施例中，处理器1404可以包括一个或多个单核或多核处理器。在一些实施例中，处理器1404可以包括通用处理器和专用处理器(例如，图形处理器，应用处理器，基带处理器等)的任意组合。在电子设备1400采用eNB(Evolved Node B，增强型基站)101或RAN(Radio Access Network，无线接入网)控制器102的实施例中，处理器1404可以被配置

为执行各种符合的实施例,例如,如图3或5所示的多个实施例中的一个或多个。

[0303] 在一些实施例中,系统控制逻辑1408可以包括任意合适的接口控制器,以向处理器1404中的至少一个和/或与系统控制逻辑1408通信的任意合适的设备或组件提供任意合适的接口。

[0304] 在一些实施例中,系统控制逻辑1408可以包括一个或多个存储器控制器,以提供连接到系统内存1412的接口。系统内存1412可以用于加载以及存储数据和/或指令。在一些实施例中电子设备1400的内存1412可以包括任意合适的易失性存储器,例如合适的动态随机存取存储器(DRAM)。

[0305] NVM/存储器1416可以包括用于存储数据和/或指令的一个或多个有形的、非暂时性的计算机可读介质。在一些实施例中,NVM/存储器1416可以包括闪存等任意合适的非易失性存储器和/或任意合适的非易失性存储设备,例如HDD(Hard Disk Drive,硬盘驱动器),CD(Compact Disc,光盘)驱动器,DVD(Digital Versatile Disc,数字通用光盘)驱动器中的至少一个。

[0306] NVM/存储器1416可以包括安装电子设备1400的装置上的一部分存储资源,或者它可以由设备访问,但不一定是设备的一部分。例如,可以经由网络接口1420通过网络访问NVM/存储1416。

[0307] 特别地,系统内存1412和NVM/存储器1416可以分别包括:指令1424的暂时副本和永久副本。指令1424可以包括:由处理器1404中的至少一个执行时导致电子设备1400实施如图3或5所示的方法的指令。在一些实施例中,指令1424、硬件、固件和/或其软件组件可另外地/替代地置于系统控制逻辑1408,网络接口1420和/或处理器1404中。

[0308] 网络接口1420可以包括收发器,用于为电子设备1400提供无线电接口,进而通过一个或多个网络与任意其他合适的设备(如前端模块,天线等)进行通信。在一些实施例中,网络接口1420可以集成于电子设备1400的其他组件。例如,网络接口1420可以集成于处理器1404的,系统内存1412,NVM/存储器1416,和具有指令的固件设备(未示出)中的至少一种,当处理器1404中的至少一个执行所述指令时,电子设备1400实现如图3或5所示的方法。

[0309] 网络接口1420可以进一步包括任意合适的硬件和/或固件,以提供多输入多输出无线电接口。例如,网络接口1420可以是网络适配器,无线网络适配器,电话调制解调器和/或无线调制解调器。

[0310] 在一个实施例中,处理器1404中的至少一个可以与用于系统控制逻辑1408的一个或多个控制器的逻辑封装在一起,以形成系统封装(SiP)。在一个实施例中,处理器1404中的至少一个可以与用于系统控制逻辑1408的一个或多个控制器的逻辑集成在同一管芯上,以形成片上系统(SoC)。

[0311] 电子设备1400可以进一步包括:输入/输出(I/O)设备1432。I/O设备1432可以包括用户界面,使得用户能够与电子设备1400进行交互;外围组件接口的设计使得外围组件也能够与电子设备1400交互。在一些实施例中,电子设备1400还包括传感器,用于确定与电子设备1400相关的环境条件和位置信息的至少一种。

[0312] 在一些实施例中,用户界面可包括但不限于显示器(例如,液晶显示器,触摸屏显示器等),扬声器,麦克风,一个或多个相机(例如,静止图像照相机和/或摄像机),手电筒(例如,发光二极管闪光灯)和键盘。

[0313] 在一些实施例中，外围组件接口可以包括但不限于非易失性存储器端口、音频插孔和电源接口。

[0314] 在一些实施例中，传感器可包括但不限于陀螺仪传感器，加速度计，近程传感器，环境光线传感器和定位单元。定位单元还可以是网络接口1420的一部分或与网络接口1420交互，以与定位网络的组件(例如，全球定位系统(GPS)卫星)进行通信。

[0315] 本申请公开的机制的各实施例可以被实现在硬件、软件、固件或这些实现方法的组合中。本申请的实施例可实现为在可编程系统上执行的计算机程序或程序代码，该可编程系统包括至少一个处理器、存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备以及至少一个输出设备。

[0316] 可将程序代码应用于输入指令，以执行本申请描述的各功能并生成输出信息。可以按已知方式将输出信息应用于一个或多个输出设备。为了本申请的目的，处理系统包括具有诸如例如数字信号处理器(DSP)、微控制器、专用集成电路(ASIC)或微处理器之类的处理器的任何系统。

[0317] 程序代码可以用高级程序化语言或面向对象的编程语言来实现，以便与处理系统通信。在需要时，也可用汇编语言或机器语言来实现程序代码。事实上，本申请中描述的机制不限于任何特定编程语言的范围。在任一情形下，该语言可以是编译语言或解释语言。

[0318] 在一些情况下，所公开的实施例可以以硬件、固件、软件或其任何组合来实现。所公开的实施例还可以被实现为由一个或多个暂时或非暂时性机器可读(例如，计算机可读)存储介质承载或存储在其上的指令，其可以由一个或多个处理器读取和执行。例如，指令可以通过网络或通过其他计算机可读介质分发。因此，机器可读介质可以包括用于以机器(例如，计算机)可读的形式存储或传输信息的任何机制，包括但不限于，软盘、光盘、光碟、只读存储器(CD-ROMs)、磁光盘、只读存储器(ROM)、随机存取存储器(RAM)、可擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、磁卡或光卡、闪存、或用于利用因特网以电、光、声或其他形式的传播信号来传输信息(例如，载波、红外信号数字信号等)的有形的机器可读存储器。因此，机器可读介质包括适合于以机器(例如，计算机)可读的形式存储或传输电子指令或信息的任何类型的机器可读介质。

[0319] 在附图中，可以以特定布置和/或顺序示出一些结构或方法特征。然而，应该理解，可能不需要这样的特定布置和/或排序。而是，在一些实施例中，这些特征可以以不同于说明性附图中所示的方式和/或顺序来布置。另外，在特定图中包括结构或方法特征并不意味着暗示在所有实施例中都需要这样的特征，并且在一些实施例中，可以不包括这些特征或者可以与其他特征组合。

[0320] 需要说明的是，本申请各设备实施例中提到的各单元/模块都是逻辑单元/模块，在物理上，一个逻辑单元/模块可以是一个物理单元/模块，也可以是一个物理单元/模块的一部分，还可以以多个物理单元/模块的组合实现，这些逻辑单元/模块本身的物理实现方式并不是最重要的，这些逻辑单元/模块所实现的功能的组合才是解决本申请所提出的技术问题的关键。此外，为了突出本申请的创新部分，本申请上述各设备实施例并没有将与解决本申请所提出的技术问题关系不太密切的单元/模块引入，这并不表明上述设备实施例并不存在其它的单元/模块。

[0321] 需要说明的是，在本专利的示例和说明书中，诸如第一和第二等之类的关系术语

仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0322] 虽然通过参照本申请的某些优选实施例，已经对本申请进行了图示和描述，但本领域的普通技术人员应该明白，可以在形式上和细节上对其作各种改变，而不偏离本申请的范围。

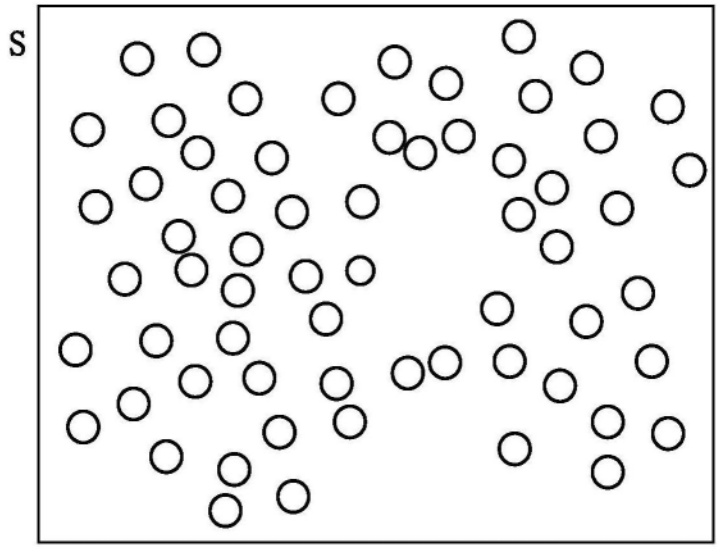


图1a

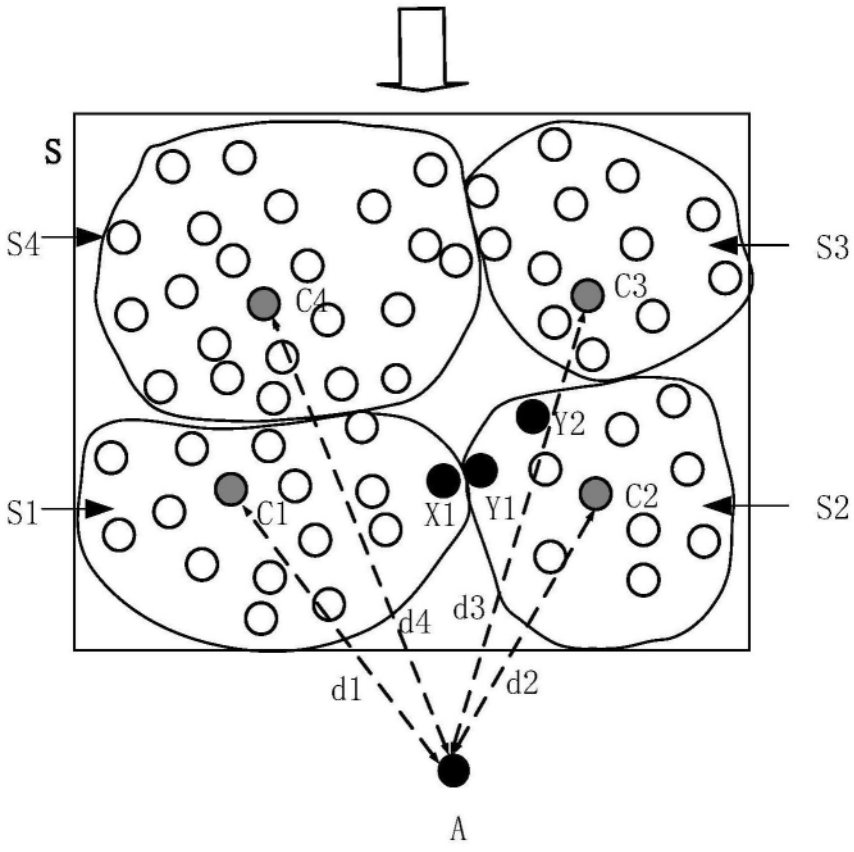


图1b

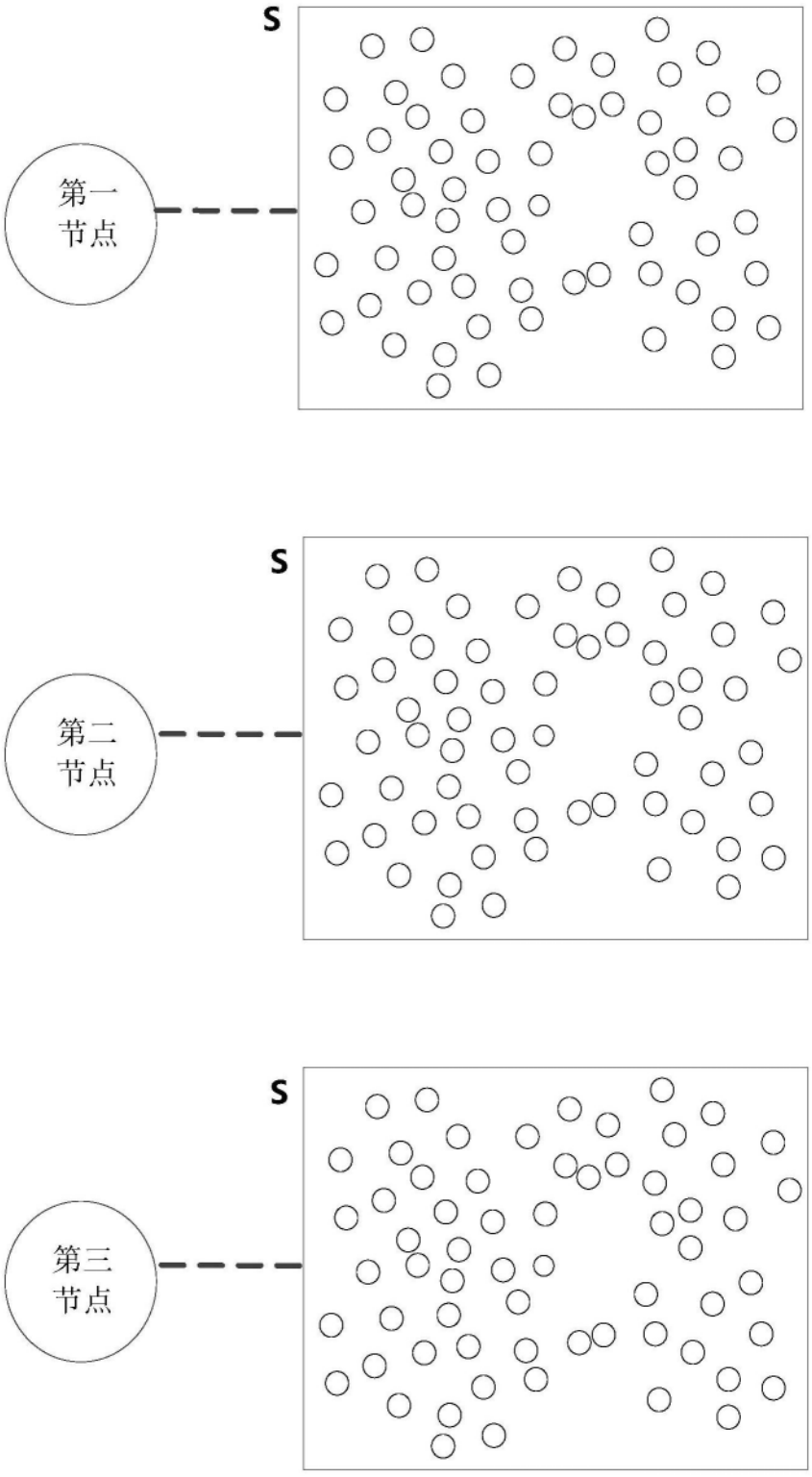


图2

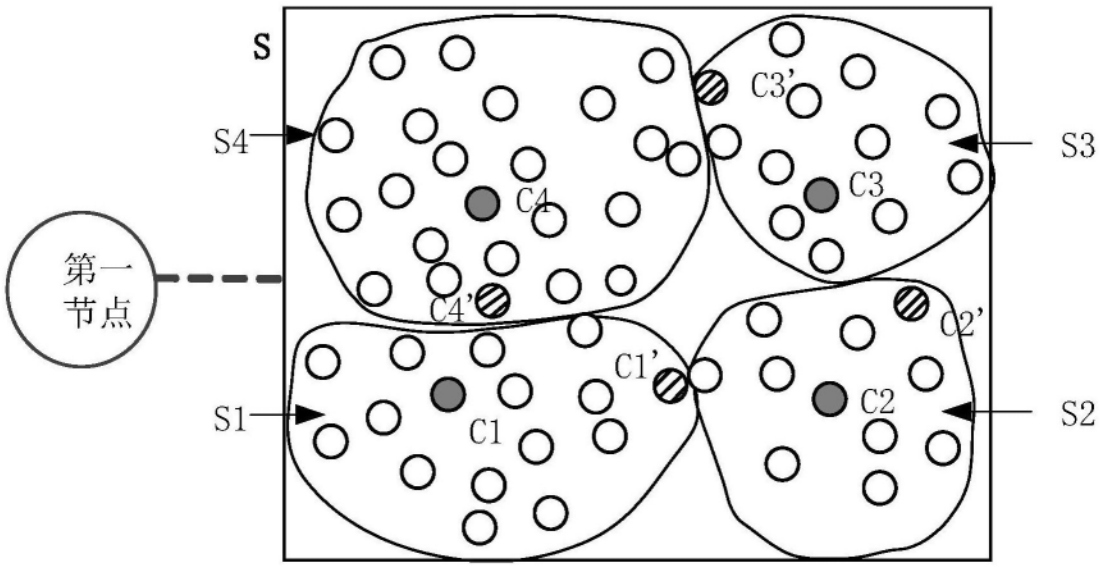

S
S4
S3
S1
S2
第一
节点
C4
C4'
C3'
C3
C2'
C1'
C1
C2

图3a

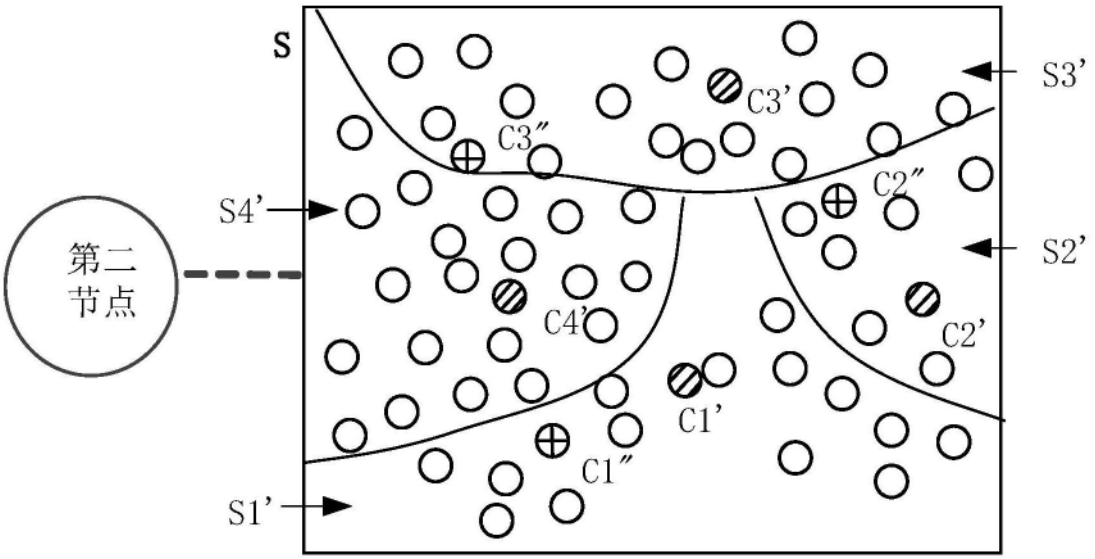

S
S3'
S4'
S2'
第二
节点
C3'
C3″
C2″
C4'
C2'
C1'
C1″
S1'

图3b

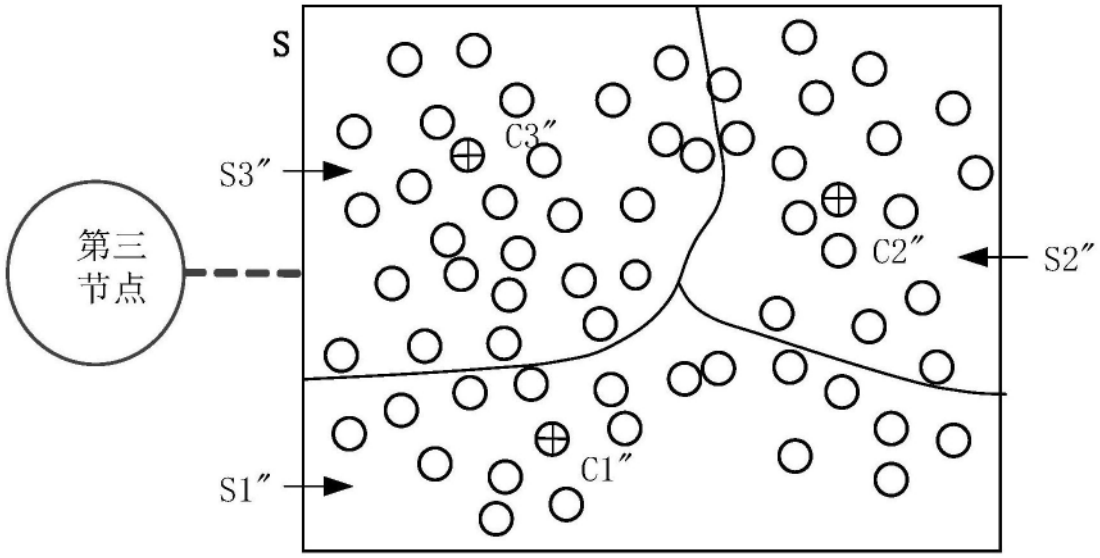


图3c

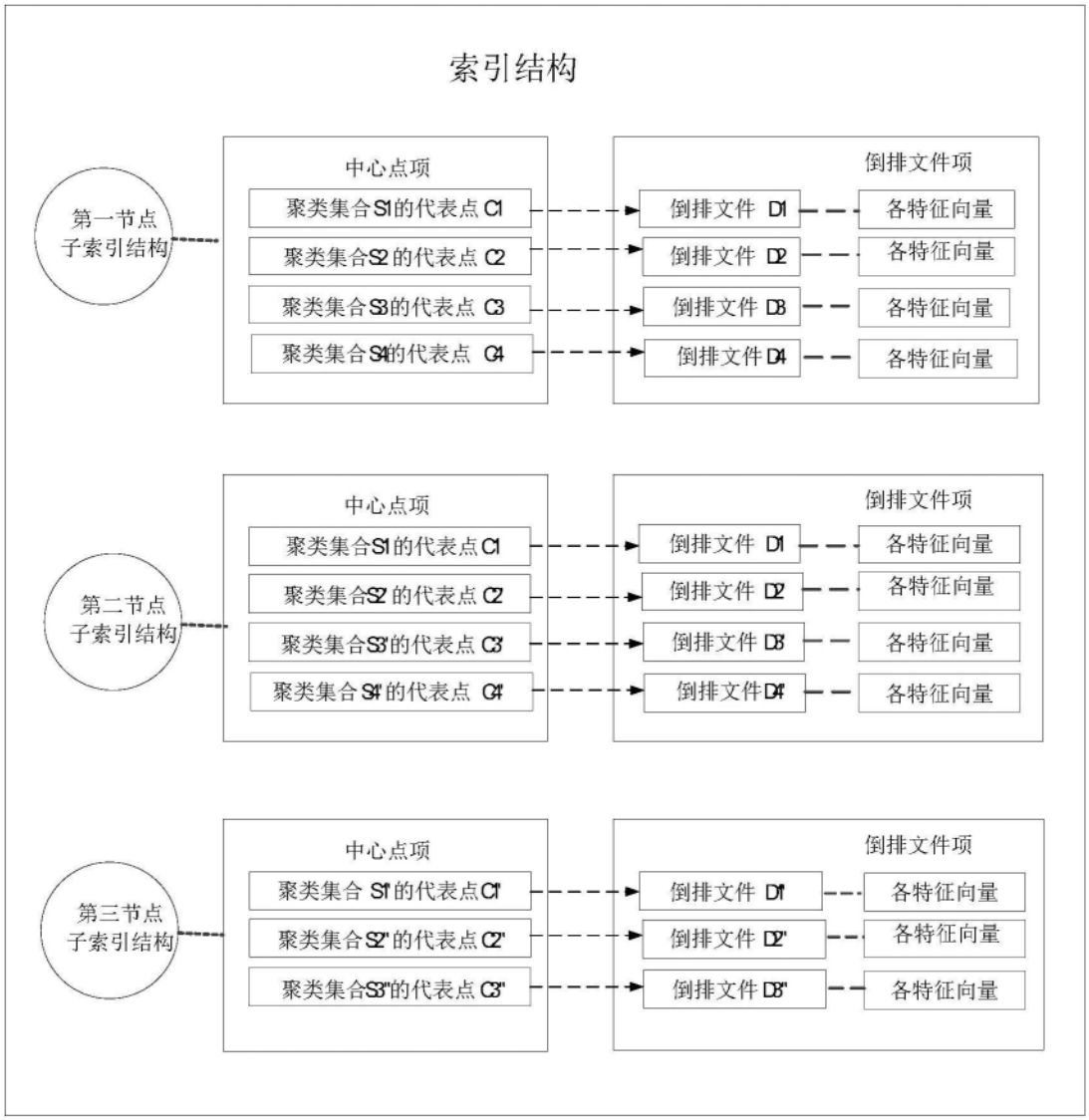

索引结构
第一节点
子索引结构
中心点项
聚类集合S1的代表点C1
聚类集合S2的代表点C2
聚类集合S3的代表点C3
聚类集合S4的代表点C4
倒排文件项
倒排文件D1
倒排文件D2
倒排文件D3
倒排文件D4
各特征向量
各特征向量
各特征向量
各特征向量
第二节点
子索引结构
中心点项
聚类集合S1的代表点C1
聚类集合S2的代表点C2
聚类集合S3'的代表点C3'
聚类集合S4'的代表点C4'
倒排文件项
倒排文件D1
倒排文件D2
倒排文件D3'
倒排文件D4'
各特征向量
各特征向量
各特征向量
各特征向量
第三节点
子索引结构
中心点项
聚类集合S1'的代表点C1'
聚类集合S2''的代表点C2''
聚类集合S3''的代表点C3''
倒排文件项
倒排文件D1'
倒排文件D2''
倒排文件D3''
各特征向量
各特征向量
各特征向量

图3d

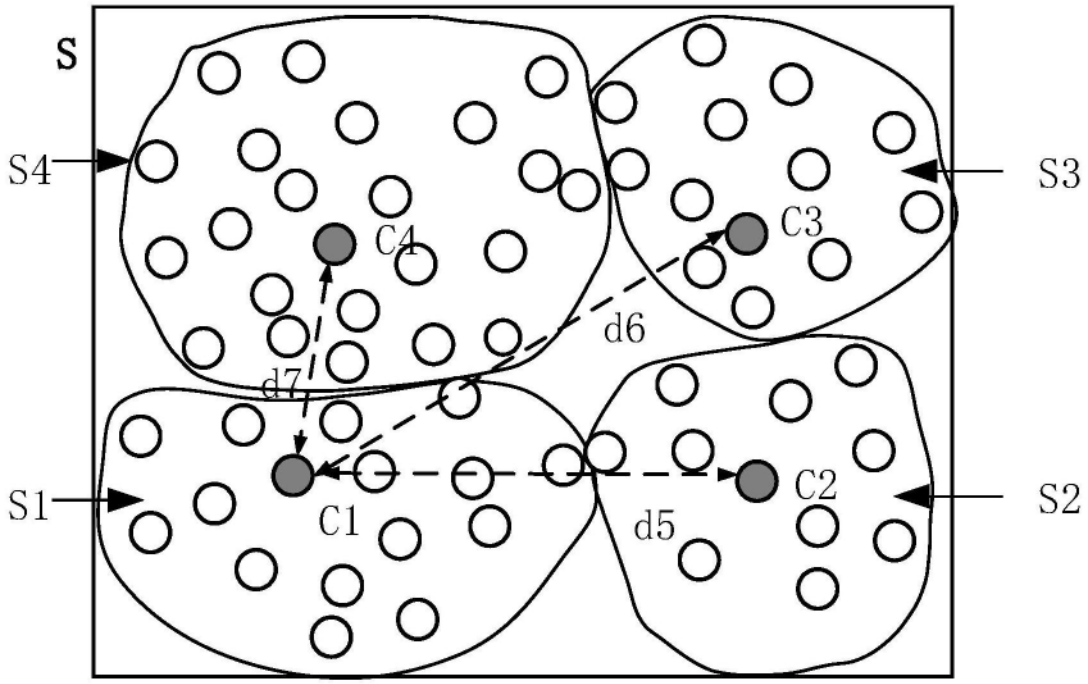

S
S4
S3
C4
C3
d6
d7
S1
C1
C2
S2
d5

图4

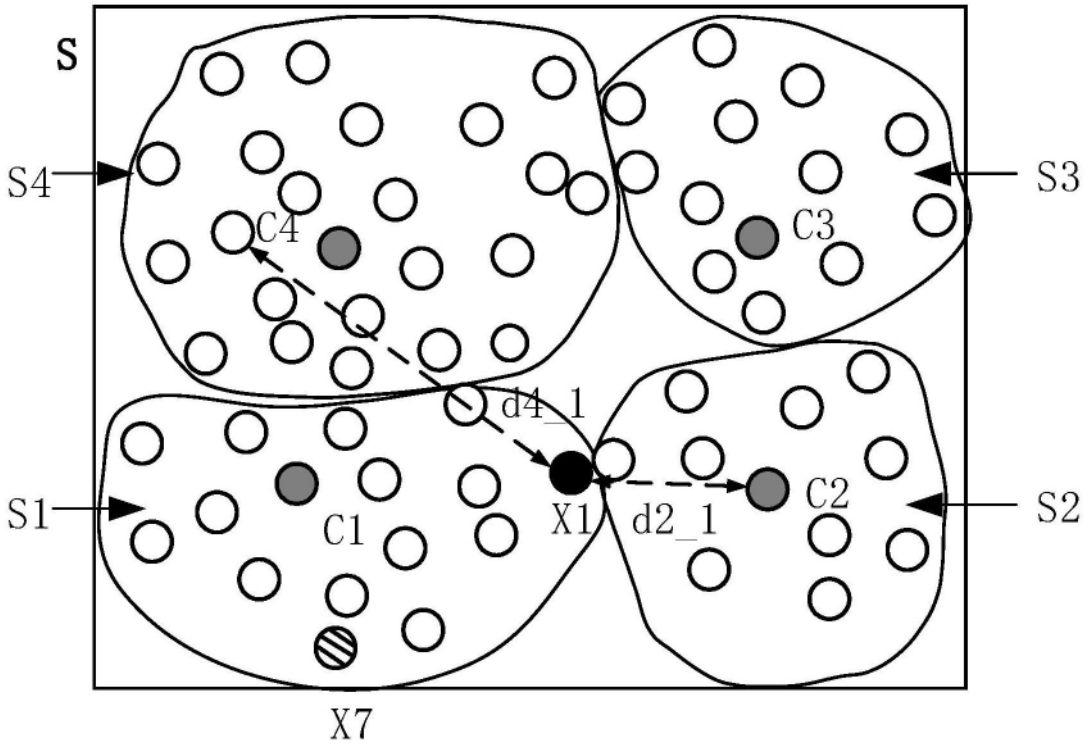

S
S4
S3
C4
C3
d4_1
S1
C1
X1
d2_1
C2
S2
X7

图5a

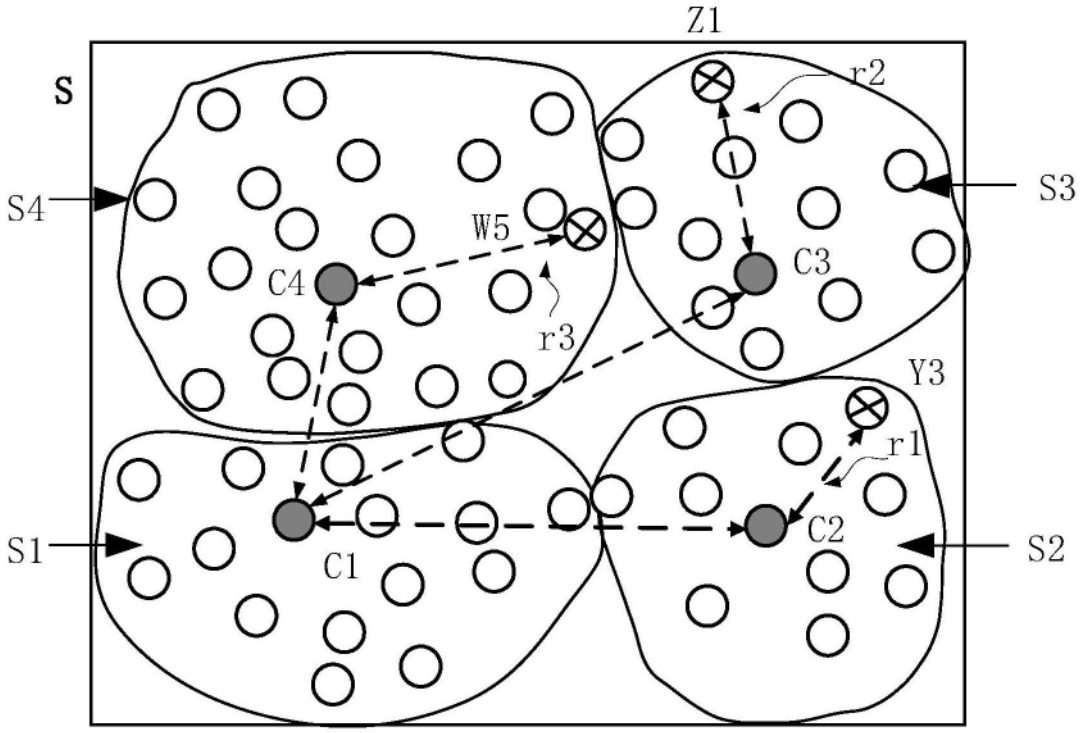

Z1
S
r2
S4
S3
W5
C4
C3
r3
Y3
r1
S1
C1
C2
S2

图5b

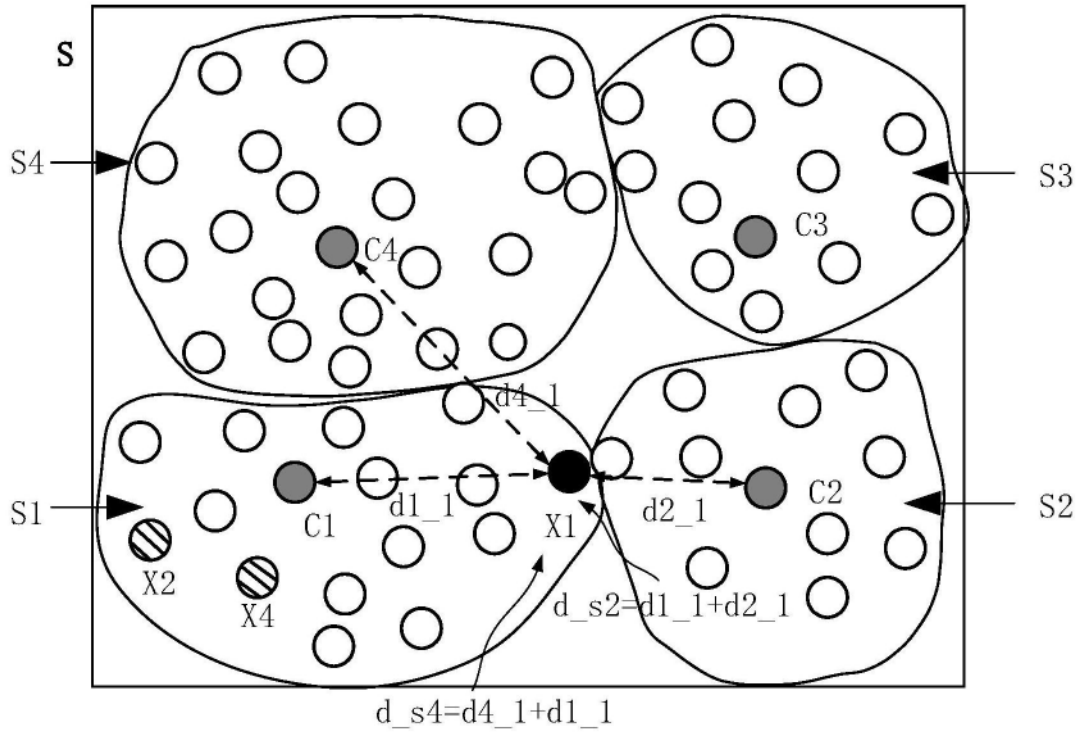

S
S4
S3
C4
C3
d4_1
S1
C1
d1_1
X1
d2_1
C2
S2
X2
X4
d_s2=d1_1+d2_1
d_s4=d4_1+d1_1

图5c

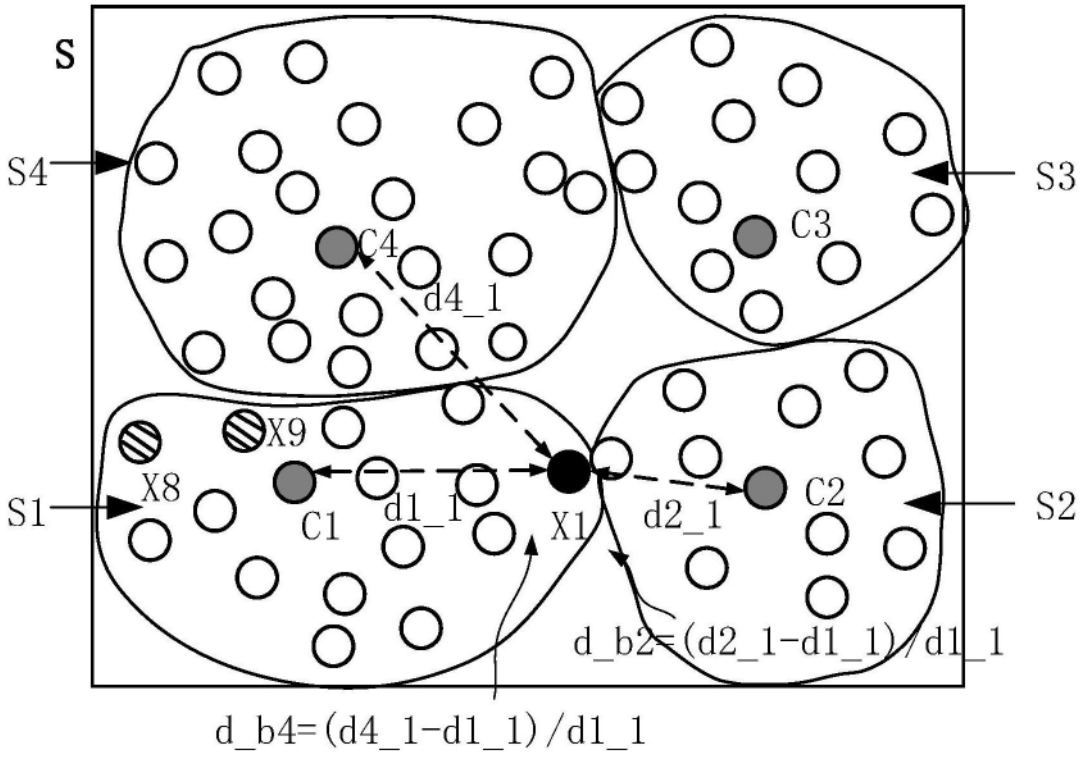

S
S4
S3
C4
C3
d4_1
X9
X8
S1
C1
d1_1
X1
d2_1
C2
S2
d_b2=(d2_1-d1_1)/d1_1
d_b4=(d4_1-d1_1)/d1_1

图5d

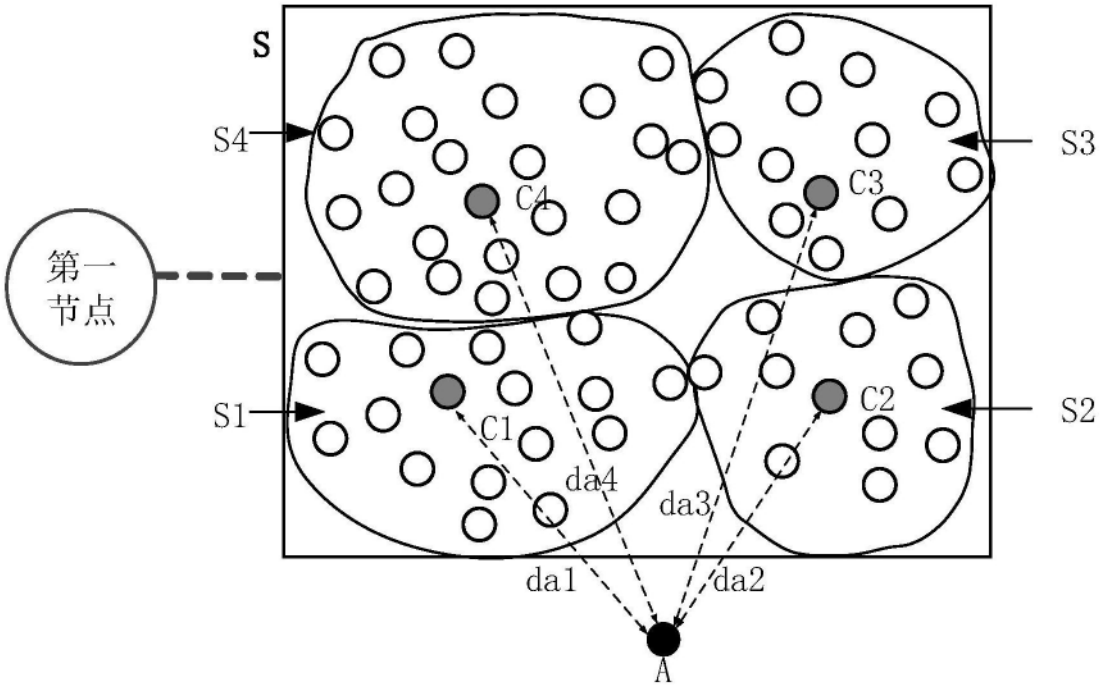

S
S4
S3
C4
C3
第一
节点
S1
C1
C2
S2
da4
da3
da1
da2
A

图6a

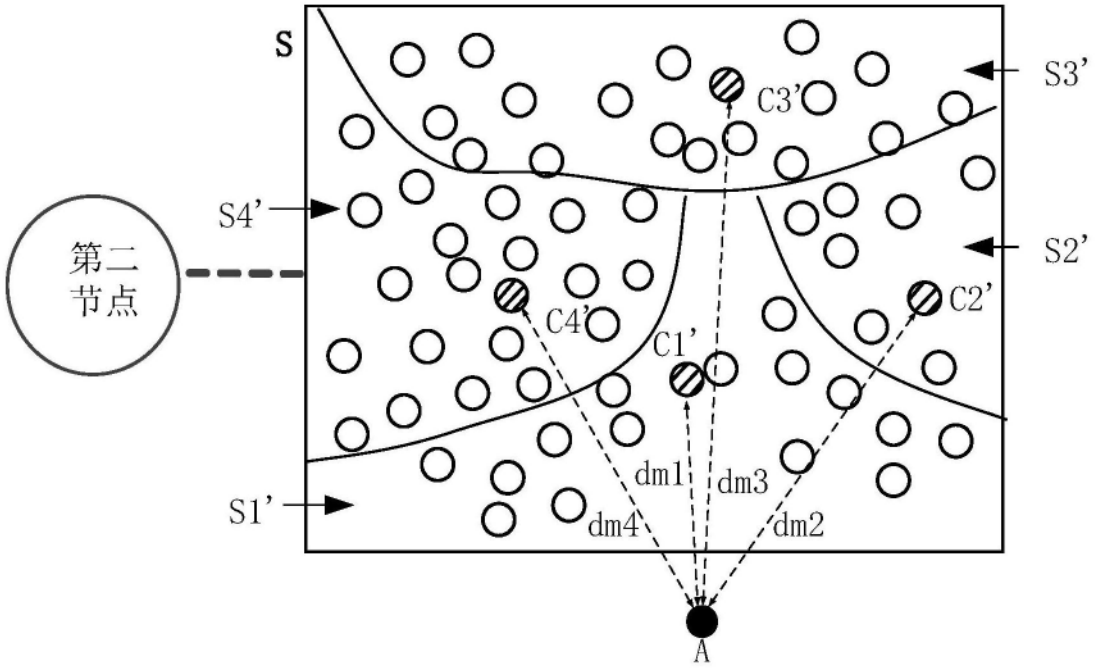

S
S3'
C3'
S4'
第二
节点
S2'
C2'
C4'
C1'
dm1
dm3
S1'
dm4
dm2
A

图6b

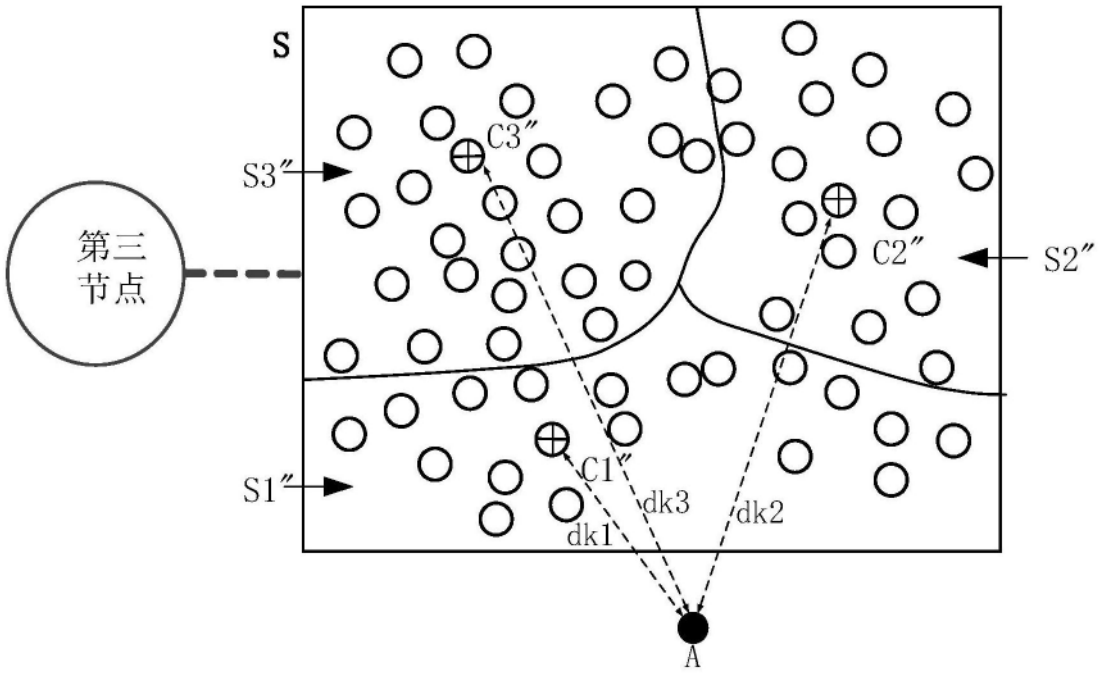

S
C3″
S3″
第三
节点
C2″
S2″
C1″
dk3
S1″
dk2
dk1
A

图6c

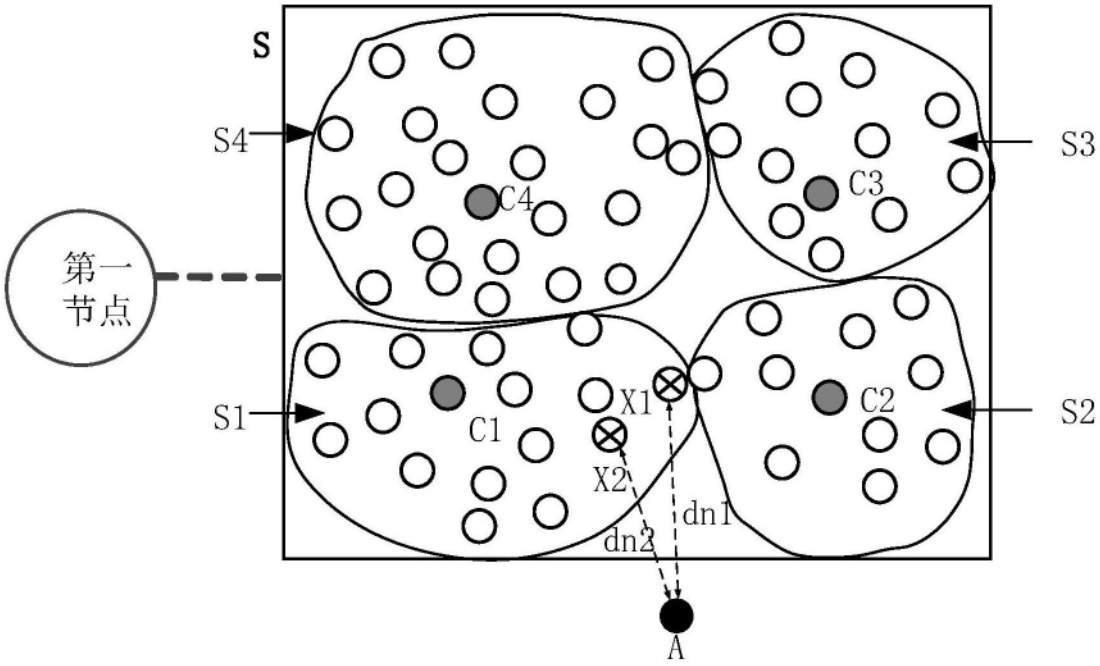

S
S4
S3
C4
C3
第一
节点
S1
C1
X1
C2
S2
X2
dn1
dn2
A

图7a

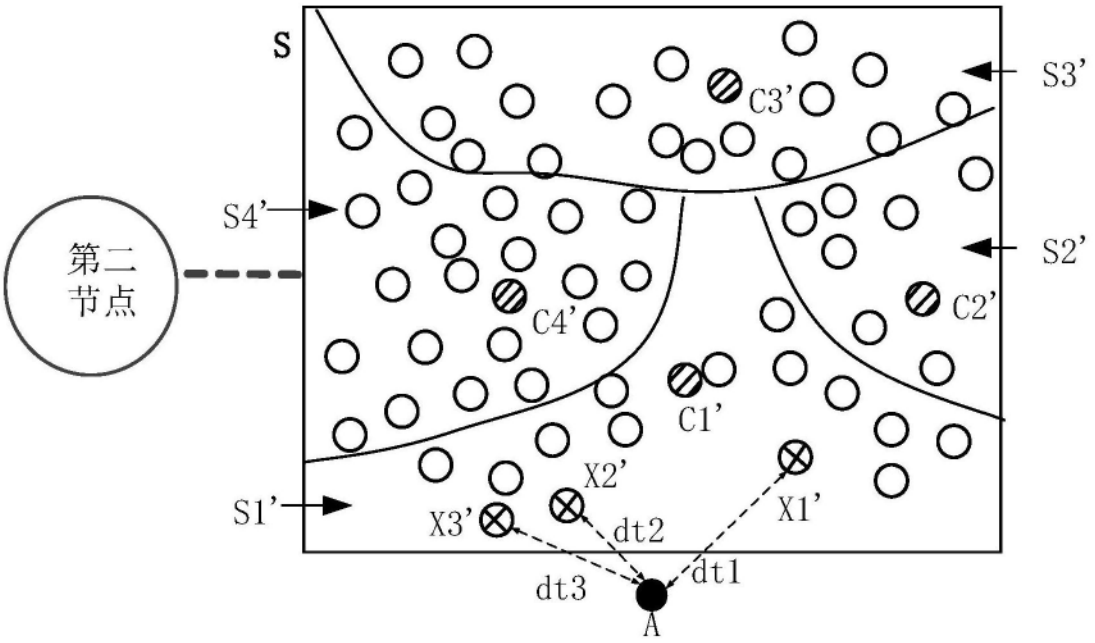

S
S3'
C3'
S4'
S2'
第二
节点
C4'
C2'
C1'
X2'
S1'
X3'
X1'
dt2
dt1
dt3
A

图7b

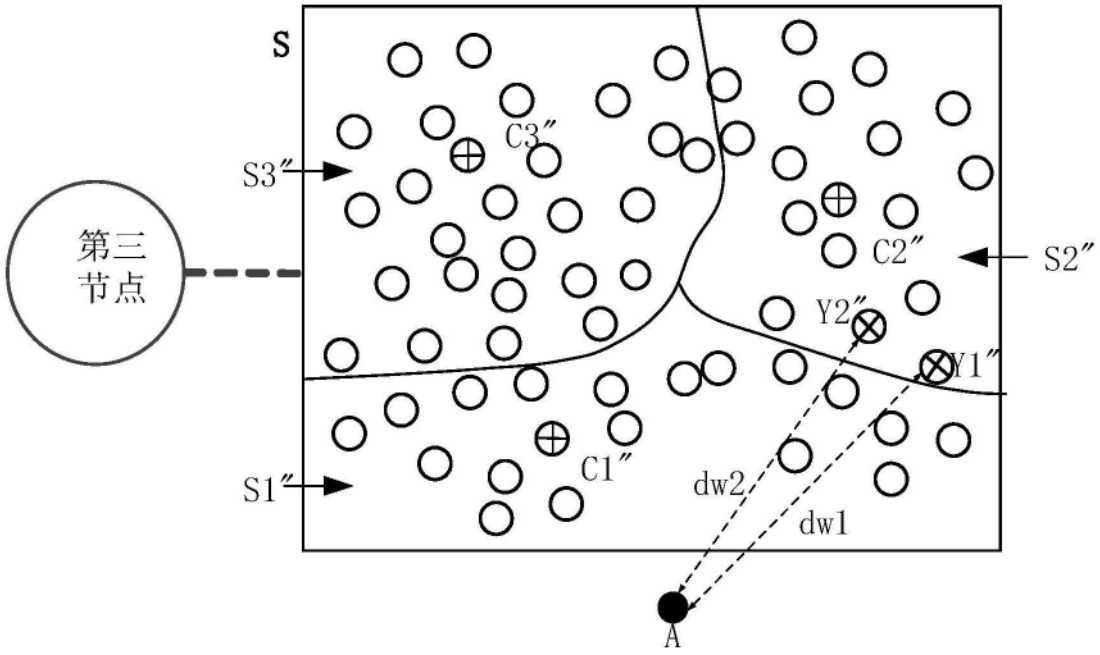


图7c

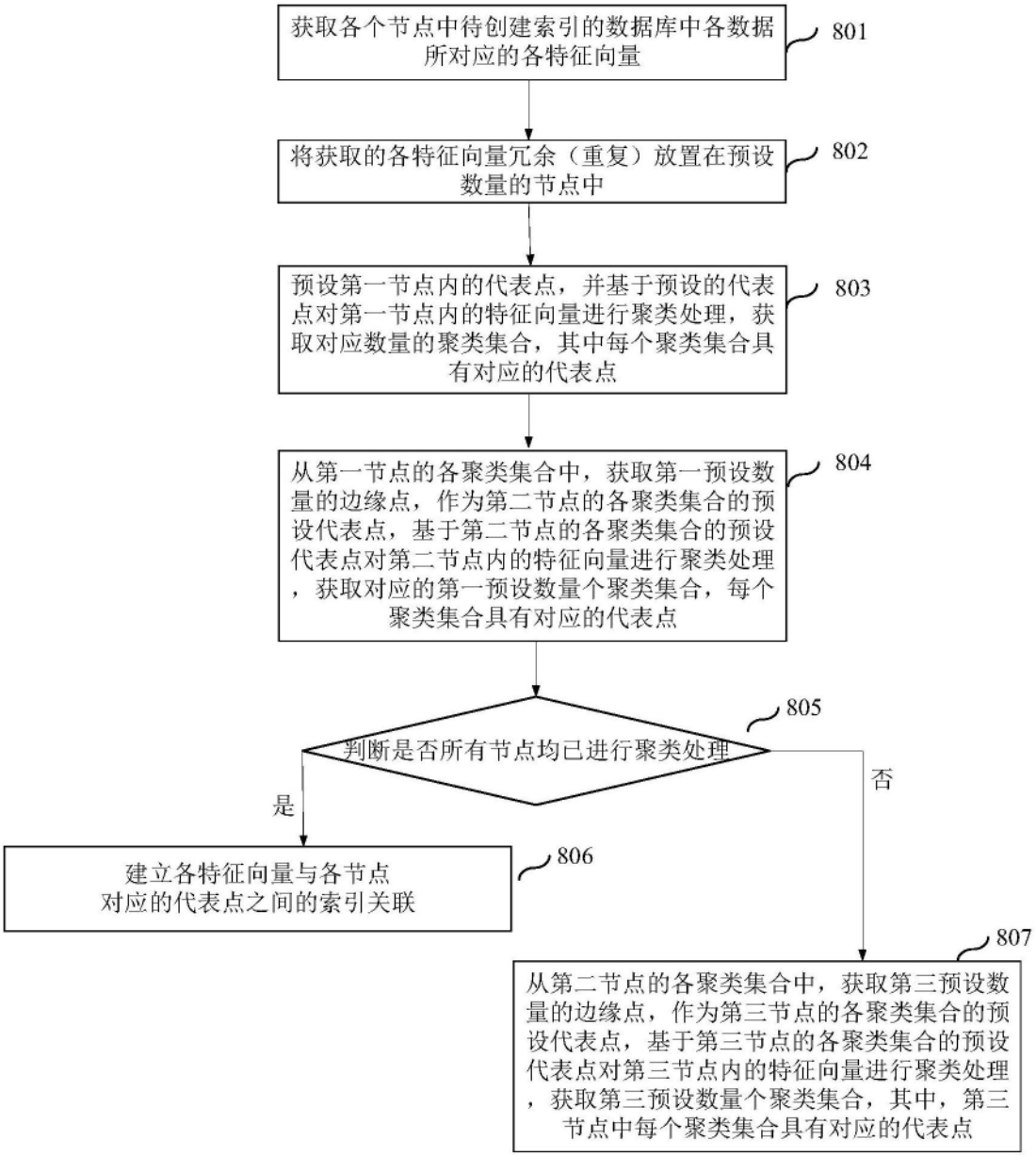

获取各个节点中待创建索引的数据库中各数据所对应的各特征向量
801
将获取的各特征向量冗余（重复）放置在预设数量的节点中
802
预设第一节点内的代表点，并基于预设的代表点对第一节点内的特征向量进行聚类处理，获取对应数量的聚类集合，其中每个聚类集合具有对应的代表点
803
从第一节点的各聚类集合中，获取第一预设数量的边缘点，作为第二节点的各聚类集合的预设代表点，基于第二节点的各聚类集合的预设代表点对第二节点内的特征向量进行聚类处理，获取对应的第一预设数量个聚类集合，每个聚类集合具有对应的代表点
804
判断是否所有节点均已进行聚类处理
805
是
否
建立各特征向量与各节点对应的代表点之间的索引关联
806
从第二节点的各聚类集合中，获取第三预设数量的边缘点，作为第三节点的各聚类集合的预设代表点，基于第三节点的各聚类集合的预设代表点对第三节点内的特征向量进行聚类处理，获取第三预设数量个聚类集合，其中，第三节点中每个聚类集合具有对应的代表点
807

图8

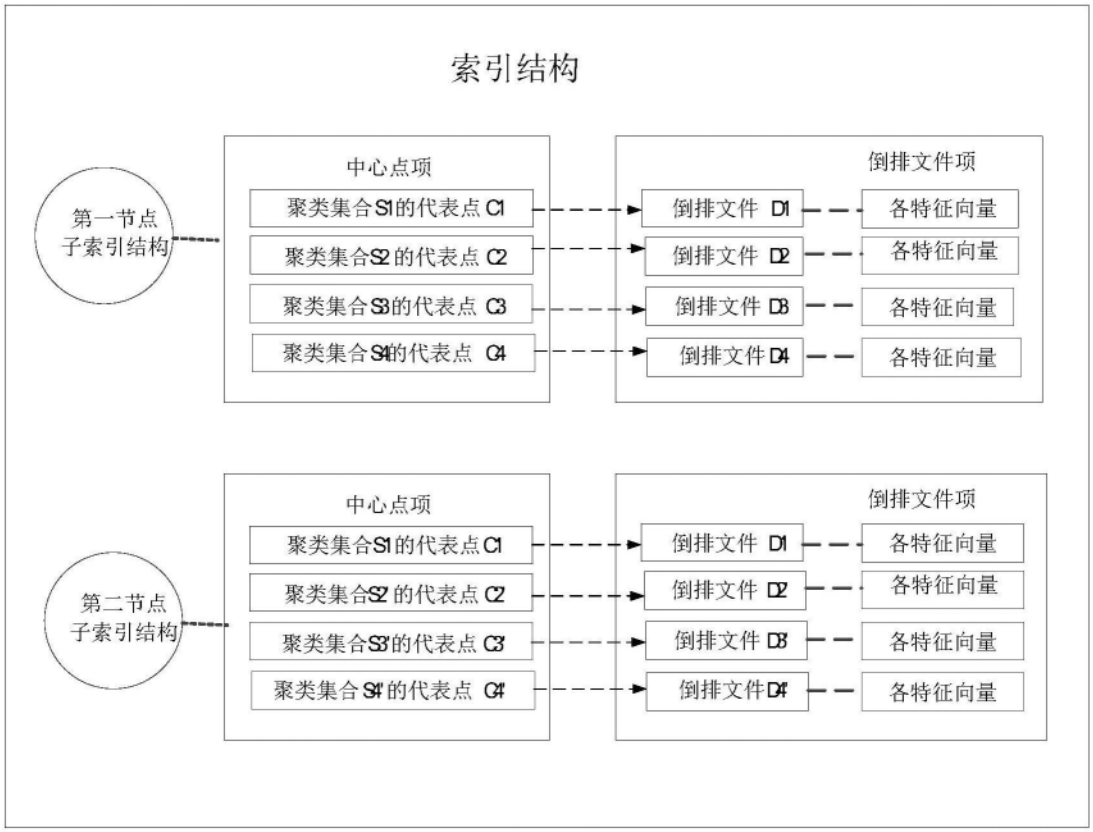

索引结构
第一节点
子索引结构
中心点项
聚类集合S1的代表点C1
聚类集合S2的代表点C2
聚类集合S3的代表点C3
聚类集合S4的代表点C4
倒排文件项
倒排文件 D1
倒排文件 D2
倒排文件 D3
倒排文件D4
各特征向量
各特征向量
各特征向量
各特征向量
第二节点
子索引结构
中心点项
聚类集合S1的代表点C1
聚类集合S2的代表点C2
聚类集合S3的代表点C3
聚类集合S4的代表点C4
倒排文件项
倒排文件 D1
倒排文件 D2
倒排文件 D3
倒排文件D4
各特征向量
各特征向量
各特征向量
各特征向量

图9

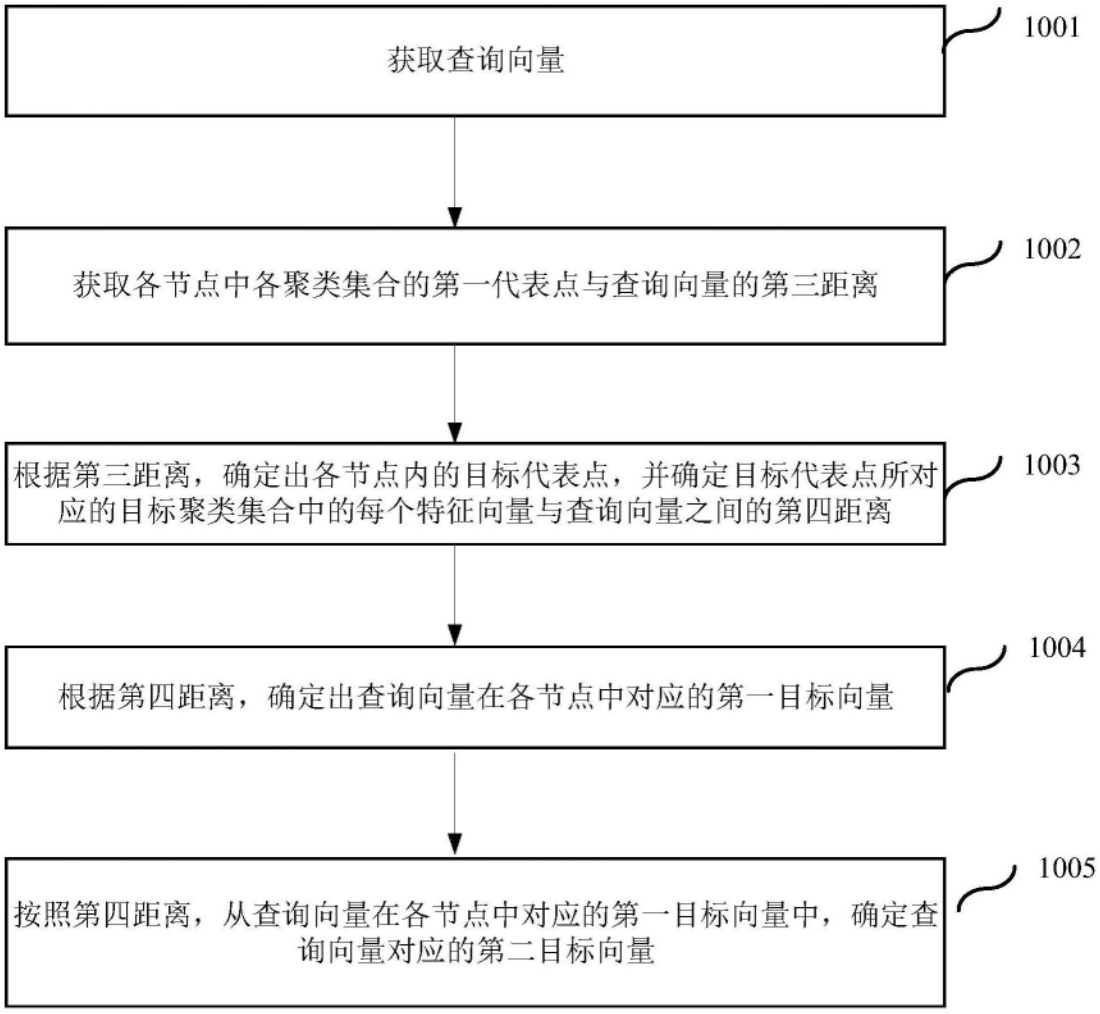

获取查询向量
1001
获取各节点中各聚类集合的第一代表点与查询向量的第三距离
1002
根据第三距离，确定出各节点内的目标代表点，并确定目标代表点所对应的目标聚类集合中的每个特征向量与查询向量之间的第四距离
1003
根据第四距离，确定出查询向量在各节点中对应的第一目标向量
1004
按照第四距离，从查询向量在各节点中对应的第一目标向量中，确定查询向量对应的第二目标向量
1005

图10

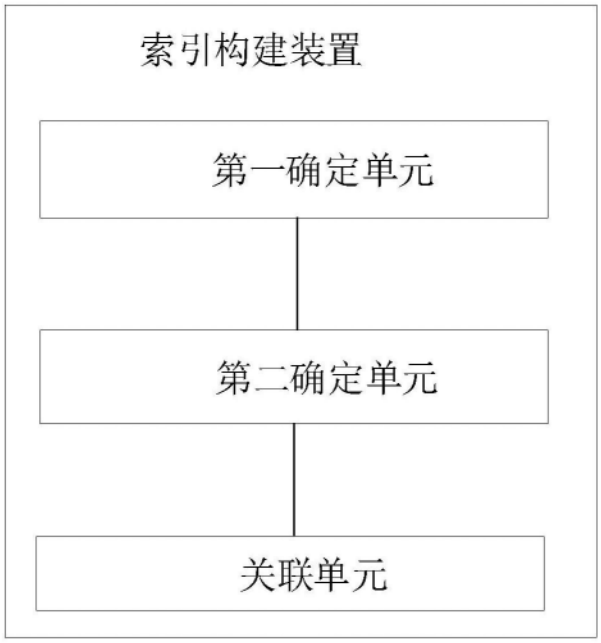

索引构建装置
第一确定单元
第二确定单元
关联单元

图11

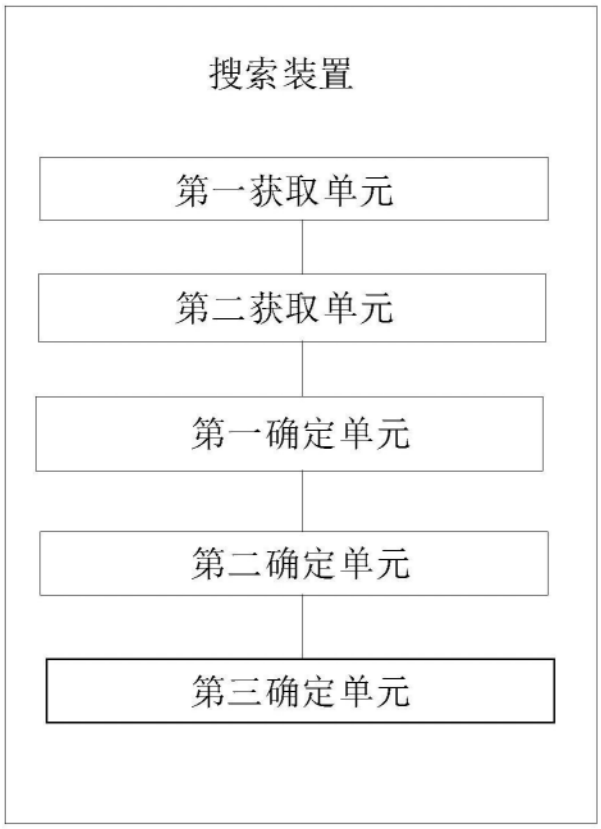

搜索装置
第一获取单元
第二获取单元
第一确定单元
第二确定单元
第三确定单元

图12

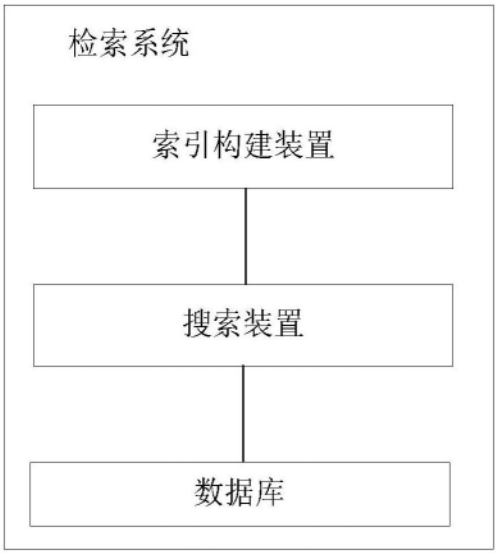

检索系统
索引构建装置
搜索装置
数据库

图13

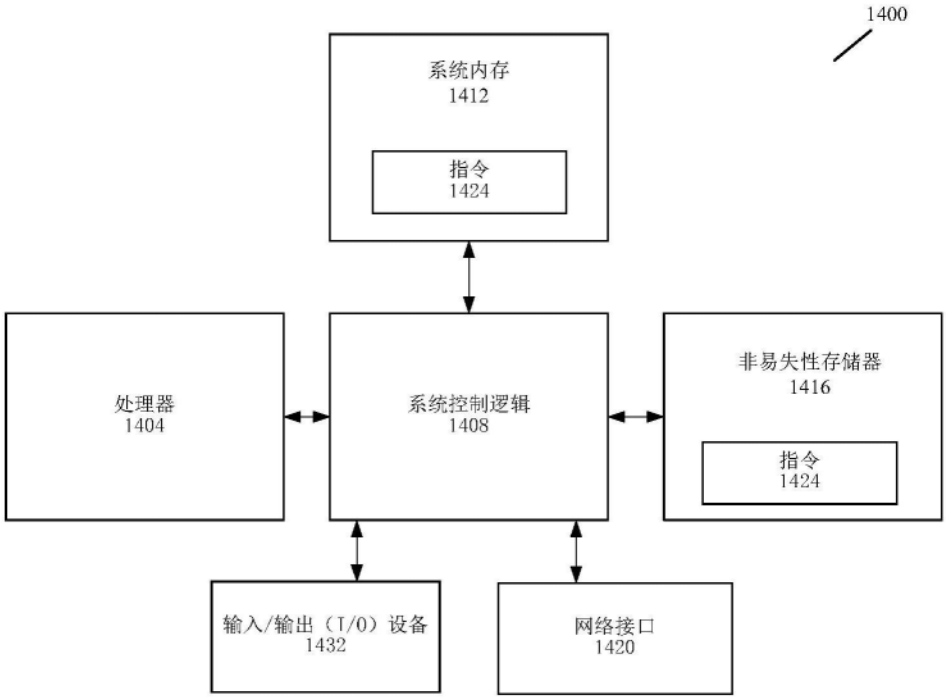

1400
系统内存
1412
指令
1424
处理器
1404
系统控制逻辑
1408
非易失性存储器
1416
指令
1424
输入/输出（I/O）设备
1432
网络接口
1420

图14