

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号
WO 2023/019933 A1

(43) 国际公布日
2023 年 2 月 23 日 (23.02.2023)

(51) 国际专利分类号:
G06F 16/21 (2019.01) **G06K 9/62** (2022.01)

(21) 国际申请号: PCT/CN2022/081154

(22) 国际申请日: 2022 年 3 月 16 日 (16.03.2022)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
202110937110.0 2021 年 8 月 16 日 (16.08.2021) CN

(71) 申请人: 百度在线网络技术(北京)有限公司 (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) [CN/CN]; 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。

(72) 发明人: 尹洁(YIN, Jie); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。 黄贲(HUANG, Ben); 中国北京市海淀区上地十街 10 号百度大厦三层, Beijing 100085 (CN)。

(74) 代理人: 北京英赛嘉华知识产权代理有限公司 (INSIGHT INTELLECTUAL PROPERTY LIMITED); 中国北京市朝阳区建外大街光华东里 8 号中海广场中楼 7 层, Beijing 100020 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW,

(54) Title: METHOD AND APPARATUS FOR CONSTRUCTING SEARCH DATABASE, AND DEVICE AND STORAGE MEDIUM

(54) 发明名称: 构建检索数据库的方法、装置、设备以及存储介质

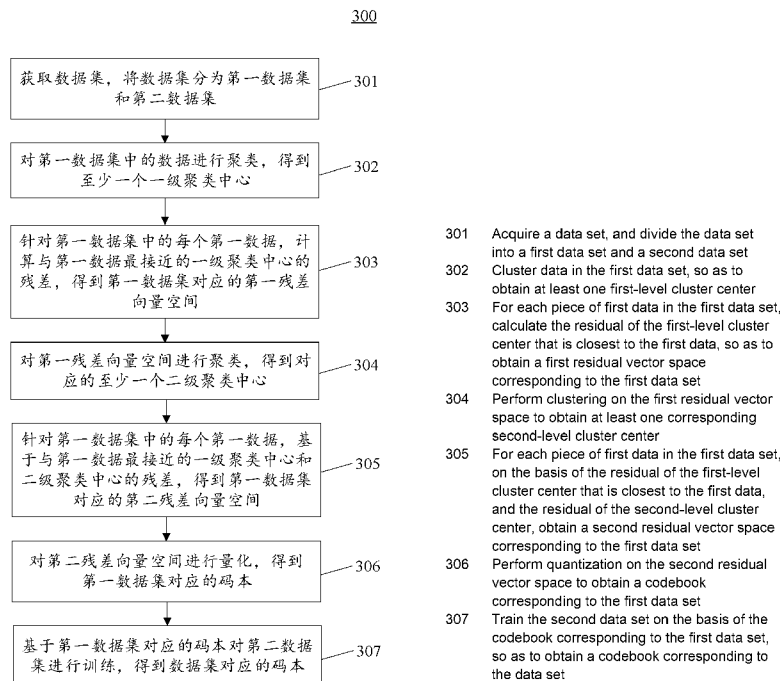


图 3

(57) Abstract: The present disclosure relates to the technical field of artificial intelligence, and particularly to the technical field of smart recommendation and deep learning. Provided are a method and apparatus for constructing a search database, and a device and a storage medium. The method comprises: acquiring a data set, and dividing the data set into a first data set and a second data set; clustering data in the first data set to obtain at least one first-level cluster center; clustering the data in the first data set on the basis of the first-level cluster center, so as to obtain at least one corresponding second-level cluster center; on the basis of the residual of

MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH,
PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK,
SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

the data in the first data set with respect to the first-level cluster center and the residual of the data in the first data set with respect to the second-level cluster center, obtaining a codebook corresponding to the first data set; and training the second data set on the basis of the codebook corresponding to the first data set, so as to obtain a codebook corresponding to the data set. By means of the method for constructing a search database in the present disclosure, a memory occupied by a search database is reduced, a search recall rate is increased, and the speed of base establishment is increased.

(57) 摘要: 本公开提供了一种构建检索数据库的方法、装置、设备以及存储介质, 涉及人工智能技术领域, 具体为智能推荐和深度学习技术领域。该方法包括: 获取数据集, 将数据集分为第一数据集和第二数据集; 对第一数据集中的数据进行聚类, 得到至少一个一级聚类中心; 基于一级聚类中心对第一数据集中的数据进行聚类, 得到对应的至少一个二级聚类中心; 基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差, 得到第一数据集对应的码本; 基于第一数据集对应的码本对第二数据集进行训练, 得到数据集对应的码本。本公开的构建检索数据库的方法, 减少了检索数据库的占用内存, 提高了检索召回率, 提高了建库的速度。

构建检索数据库的方法、装置、设备以及存储介质

交叉引用

本专利申请要求于 2021 年 08 月 16 日提交的、申请号为 202110937110.0、发明名称为“构建检索数据库的方法、装置、设备以及存储介质”的中国专利申请的优先权，该申请的全文以引用的方式并入本申请中。

技术领域

本公开涉及人工智能技术领域，具体为智能推荐和深度学习技术领域，尤其涉及构建检索数据库的方法、装置、设备以及存储介质。

背景技术

随着 CNN (Convolutional Neural Network, 卷积神经网络) 的广泛应用，使得应用在 CNN 特征上的基础检索技术 ANN (Approximate Nearest Neighbor, 近似最近邻检索) 也得到了较快发展。针对大规模数据集，当前业内主流的检索算法有 HNSW (Hierarchical Navigable Small World)、IVF (Inverted File System) 等。但是 HNSW 算法的计算量大，内存需求大；IVF 算法中的聚类中心数目多，迭代次数多，整体计算量接近指数级增长。

发明内容

本公开提供了一种构建检索数据库的方法、装置、设备以及存储介质。

根据本公开的第一方面，提供了一种构建检索数据库的方法，包括：获取数据集，将数据集分为第一数据集和第二数据集，其中，第一数据集中的数据量小于第二数据集中的数据量；对第一数据集中的数据进行聚类，得到至少一个一级聚类中心；基于一级聚类中心对第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差，得到第一数据集对应的码本；基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。

根据本公开的第二方面，提供了一种检索方法，包括：获取待检索数据；计算待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心，其中，检索数据库通过如第一方面中任一实现方式描述的方法得到；计算待检索数据与第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；基于检索数据库中的数据集对应的码本，依次获取第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；计算待检索数据与样本数据的距离，基于计算结果返回第五预设数目个样本数据。

根据本公开的第三方面，提供了一种构建检索数据库的装置，包括：第一获取模块，被配置成获取数据集，将数据集分为第一数据集和第二数据集，其中，第一数据集中的数据量小于第二数据集中的数据量；第一聚类模块，被配置成对第一数据集中的数据进行聚类，得到至少一个一级聚类中心；第二聚类模块，被配置成基于一级聚类中心对第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；得到模块，被配置成基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差，得到第一数据集对应的码本；训练模块，被配置成基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。

根据本公开的第四方面，提供了一种检索装置，包括：第二获取模块，被配置成获取待检索数据；第一计算模块，被配置成计算待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心，其中，检索数据库通过如第一方面中任一实现方式描述的方法得到；第二计算模块，被配置成计算待检索数据与第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；第三获取模块，被配置成基于检索数据库中的数据集中的数据集对应的码本，依次获取第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；第三计算模块，被配置成计算待检索数据与样本数据的距离，基于计算结果返回第五预设数目个样本数据。

根据本公开的第五方面，提供了一种电子设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；其中，存储器存储有可被至少一个处理器执行的指令，指令被至少一个处理器执行，以使至少一个处理器能

够执行如第一方面或第二方面中任一实现方式描述的方法。

根据本公开的第六方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，计算机指令用于使计算机执行如第一方面或第二方面中任一实现方式描述的方法。

根据本公开的第七方面，提供了一种计算机程序产品，包括计算机程序，计算机程序在被处理器执行时实现如第一方面或第二方面中任一实现方式描述的方法。

应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明而变得容易理解。

附图说明

附图用于更好地理解本方案，不构成对本公开的限定。其中：

图 1 是本公开可以应用于其中的示例性系统架构图；

图 2 是根据本公开的构建检索数据库的方法的一个实施例的流程图；

图 3 是根据本公开的构建检索数据库的方法的另一个实施例的流程图；

图 4 是图 3 所示的构建检索数据库的方法的训练步骤的分解流程图；

图 5 是根据本公开的检索方法的一个实施例的流程图；

图 6 是根据本公开的构建检索数据库的装置的一个实施例的结构示意图；

图 7 是根据本公开的检索装置的一个实施例的结构示意图；

图 8 是用来实现本公开实施例的构建检索数据库的方法的电子设备的框图。

具体实施方式

以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特

征可以相互组合。下面将参考附图并结合实施例来详细说明本公开。

图 1 示出了可以应用本公开的构建检索数据库的方法或构建检索数据库的装置的实施例的示例性系统架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 101、102、103 通过网络 104 与服务器 105 交互，以接收或发送信息等。终端设备 101、102、103 上可以安装有各种客户端应用。

终端设备 101、102、103 可以是硬件，也可以是软件。当终端设备 101、102、103 为硬件时，可以是各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。当终端设备 101、102、103 为软件时，可以安装在上述电子设备中。其可以实现成多个软件或软件模块，也可以实现成单个软件或软件模块。在此不做具体限定。

服务器 105 可以提供各种服务。例如，服务器 105 可以对从终端设备 101、102、103 获取到的数据集进行分析和处理，并生成处理结果（例如数据集对应的码本）。

需要说明的是，服务器 105 可以是硬件，也可以是软件。当服务器 105 为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器 105 为软件时，可以实现成多个软件或软件模块（例如用来提供分布式服务），也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本公开实施例所提供的构建检索数据库的方法一般由服务器 105 执行，相应地，构建检索数据库的装置一般设置于服务器 105 中。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

继续参考图 2，其示出了根据本公开的构建检索数据库的方法的一个实施例的流程 200。该构建检索数据库的方法包括以下步骤：

步骤 201，获取数据集，将数据集分为第一数据集和第二数据集。

在本实施例中，构建检索数据库的方法的执行主体（例如图 1 所示的服务器 105）可以获取数据集。该数据集中可以包括通过各种方式获取的原始数据，原始数据中的数据可以为图像、视频和/或音频等，然后提取每个原始数据的特征，从而得到原始数据集对应的特征集合，也即上述数据集。数据集的数据量一般可以达到千亿甚至万亿级别，且其中的每个数据一般都是多维度的。

上述执行主体在获取到数据集后，会将数据集分为第一数据集和第二数据集，例如，可通过对数据集进行随时抽样，从而得到第一数据集，然后将数据集中除去第一数据集的部分作为第二数据集。需要说明的是，第一数据集中的数据量要远小于第二数据集中的数据量，例如，当数据集中的数据量为 1 亿时，第一数据集中的数据量可为 500 万。

步骤 202，对第一数据集中的数据进行聚类，得到至少一个一级聚类中心。

在本实施例中，上述执行主体可以利用聚类算法对第一数据集中的数据进行聚类，从而得到至少一个一级聚类中心。聚类算法是指将一堆没有标签的数据自动划分成几类的方法，属于无监督学习方法，其是根据样本之间的距离或者说是相似性（亲疏性），把越相似、差异越小的样本聚成一类（簇），最后形成多个簇，并可以得到每个簇的中心，也即聚类中心，得到的聚类中心与其覆盖范围内的所有数据有关，为所有数据的均值向量。聚类中心向量能全面描述其下的数据分布，且不同聚类中心之间有一定的区分度。

可选地，可使用 K-Means（K 均值）聚类算法对第一数据集中的数据进行聚类，得到对应的一级子空间以及对应的至少一个一级聚类中心。具体地，先设置 K 值，也即设置希望得到的簇的个数，然后从第一数据集中随机选择 K 个数据点作为质心（聚类中心），之后对第一数据集中每一个点，计算其与每一个质心的距离（如欧式距离等），离哪个质心近，就划分到那个质心所属的集合。把所有数据归好集合后，一共有 K 个集合，然后重新计算每个集合的质心，如果新计算出来的质心和原来的质心之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定或收敛），则可以认为聚类已经达到期望的结果，算法终止，但是如果新质心和原质心距离变化很大，需要迭代上述步骤，直至趋于稳定。当然，也可选择其他的聚类算法对第一数据集中的数据进行聚类，本公开对此不做具体限定。

在一级子空间内的数据后续会再次被分组,所以在本步骤中无需对一级子空间内的数据进行精细划分,保证不同一级聚类中心有一定的区分度即可。

步骤 203, 基于一级聚类中心对第一数据集中的数据进行聚类, 得到对应的至少一个二级聚类中心。

在本实施例中, 上述执行主体可以基于步骤 202 得到的一级聚类中心对第一数据集中的数据再次进行聚类, 从而得到对应的二级子空间以及对应的至少一个二级聚类中心。其中, 二级子空间内的数据与一级子空间内的数据分布是相似的, 且二级聚类中心中的多个子集能够准确的描述每个子空间。

作为示例, 可计算第一数据集中的第一数据与其最接近的一级聚类中心的残差向量, 从而得到每个第一数据对应的残差向量, 上述每个第一数据对应的残差向量构成残差向量空间, 残差向量空间描述了数据与聚类中心的相对位置分布。然后使用 K-Means 聚类算法对该残差向量空间进行聚类, 以描述新的空间内的数据分布, 从而得到对应的至少一个二级聚类中心。

需要说明的是, 在一级子空间内, 属于同一个聚类中心下的数据, 在二级子空间可以属于任意一个聚类中心; 在二级子空间内属于同一个聚类中心的残差向量, 可能来自不同的一级子空间的聚类中心。而在一级子空间与二级子空间内, 均属于同一个聚类中心的数据, 则认为是高度相似的。

例如, 假设在数据量为 4 亿规模的数据集上, 一级子空间 A 个数为 5000 时, 平均每个子空间下样本总数约为 80000, 二级子空间 B 再划分为 5000 个子空间。对于 A 中某个子空间 C_i , 其下所有样本允许属于 $\{F_{i1}, F_{i2}, \dots, F_{it}\}$, $\{F_{i1}, F_{i2}, \dots, F_{it}\}$ 是 $\{F_1, F_2, \dots, F_t\}$ 的一个子集; 每个样本均可以用聚类中心对标识, 即 $\{(C_i, F_{i1}), (C_i, F_{i2}), \dots, (C_i, F_{it})\}$ 。此时, 训练成本为 $2*5000$ 个聚类中心, 相比 $200000 (\sqrt{n}*c)$ 成本大幅度降低。另外, 通过对聚类中心进行分层, 减少了在建库过程中的计算量, 从而提高了建库效率和检索的效率。

步骤 204, 基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差, 得到第一数据集对应的码本。

在本实施例中, 上述执行主体可以基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差, 得到第一数据集对应的码本。作为示例, 可计算第一数据集中的第一数据所属于的一级聚类中心以及二级聚类中心, 然后计算第一数据集中的第一数据与一级聚类中心以及二级聚类中心的残差, 从而得到一个新的特征空间(残差向量空间), 然后在这个新的特征空间内进

行量化，由于新的特征空间中的特征的取值范围大幅度缩小，所以量化损失也可大幅度缩小。

可选地，量化方式可为 PQ 量化（Product Quantization，乘积量化），PQ 量化属于一种经典的 ANN 算法，PQ 量化可减少存储、减少运算量以及减少运算量。对新的特征空间进行 PQ 量化，可得到第一数据集对应的码本。在使用 PQ 量化时，一般使用四分之一量化。

步骤 205，基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。

在本实施例中，上述执行主体可以基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。也即上述执行主体可以用第一数据集对应的码本中的聚类中心的 ID（Identity Document，身份标识号）标识第二数据集中的特征向量，从而在检索过程中，无需加载原始特征向量，降低了检索过程中所需要的内存。

本公开实施例提供的构建检索数据库的方法，首先获取数据集，将数据集分为第一数据集和第二数据集；然后对第一数据集中的数据进行聚类，得到至少一个一级聚类中心；之后基于一级聚类中心对第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；并基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差，得到第一数据集对应的码本；最后基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。本实施例中的构建检索数据库的方法属于倒排索引法，该方法通过将特征向量空间进行空间分割，缩短了训练时间，使得到的聚类中心能够更好地描述数据分布；此外，通过对新的特征空间进行量化，降低了量化损失，并且降低了检索过程中所需要的内存；并且，该方法支持多种建库模式，支持实时/日/周/月级别的索引更新，从而提供了高召回、高吞吐、低时延和低内存的基础检索能力。

本公开的技术方案中，所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理，均符合相关法律法规的规定，且不违背公序良俗。

继续参考图 3，图 3 示出了根据本公开的构建检索数据库的方法的另一个实施例的流程 300。该构建检索数据库的方法包括以下步骤：

步骤 301, 获取数据集, 将数据集分为第一数据集和第二数据集。

步骤 302, 对第一数据集中的数据进行聚类, 得到至少一个一级聚类中心。

步骤 301-302 与前述实施例的步骤 201-202 基本一致, 具体实现方式可以参考前述对步骤 201-202 的描述, 此处不再赘述。

步骤 303, 针对第一数据集中的每个第一数据, 计算与第一数据最接近的一级聚类中心的残差, 得到第一数据集对应的第一残差向量空间。

在本实施例中, 针对第一数据集中的每个第一数据, 构建检索数据库的方法的执行主体 (例如图 1 所示的服务器 105) 可以计算与其最接近的一级聚类中心的残差, 从而得到第一数据集对应的第一残差向量空间, 该残差向量空间描述了数据与聚类中心的相对位置分布。其中, 与其最接近的一级聚类中心可通过距离计算结果来确定。

步骤 304, 对第一残差向量空间进行聚类, 得到对应的至少一个二级聚类中心。

在本实施例中, 上述执行主体可以对步骤 303 得到的第一残差向量空间进行聚类, 从而得到对应的至少一个二级聚类中心。具体地, 可采用 K-Means 聚类算法对第一残差向量空间进行聚类。一二级聚类中心将整个数据空间分割为多个子空间, 且所有一级聚类中心共享二级聚类中心。

通过对第一数据集中的数据进行聚类, 得到一级聚类中心以及二级聚类中心, 从而对大规模的数据集进行了更精细的空间分割, 使得聚类中心向量能更好地代表其下所有数据, 进而更好地描述数据的分布情况。

步骤 305, 针对第一数据集中的每个第一数据, 基于与第一数据最接近的一级聚类中心和二级聚类中心的残差, 得到第一数据集对应的第二残差向量空间。

在本实施例中, 针对第一数据集中的每个第一数据, 上述执行主体可以基于与每个第一数据最接近的一级聚类中心和二级聚类中心的残差, 得到第一数据集对应的第二残差向量空间。例如, 可计算与每个第一数据最接近的一级聚类中心和二级聚类中心的残差, 从而得到对应的第二残差向量空间。

步骤 306, 对第二残差向量空间进行量化, 得到第一数据集对应的码本。

在本实施例中, 上述执行主体可以对步骤 305 得到的第二残差向量空间进行量化, 从而得到第一数据集对应的码本。可选地, 量化方式可为 PQ 量

化，一般使用四分之一量化，PQ 量化损失在每个特征维度内表现为特征值与聚类中心的差值，当样本特征在每个维度的取值范围较大时，用固定个聚类中心表示必然导致量化损失增大，故本步骤对第二残差向量空间进行量化，从而缩小特征的取值范围，进而缩小量化损失。PQ 量化可减少存储、减少运算量以及减少运算量。

在本实施例的一些可选实施方式中，步骤 306 包括：将第二残差向量空间分为第一预设数目个子空间；分别对第一预设数目个子空间中的每个子空间进行聚类，得到第二预设数目个空间聚类中心；分别对空间聚类中心的标识号 ID 进行标记，得到每个子空间的码本；将每个子空间的码本做直积，得到第一数据集对应的码本。其中，第一预设数目与第二预设数目可根据实际情况进行设定，本公开对此不做具体限定。

具体地，假设有 N 个训练样本，每个样本维度 D 是 64。把每个样本分割为 $M=8$ 段，对每一段聚类，聚为 $K=256$ 类，得到 $M*K$ 个聚类中心，每个聚类中心维度为 D/M ，这些聚类中心称为码本，训练得到码本，并保存。对于每个子段 M ，去 K 个聚类中心中寻找距离最近的聚类中心 $U(M)$ 。存储每一个类对应子类的标签列表。这样训练完成，对 N 个样本，最后存下来的仅有 $K*M$ 个聚类中心和 $N*M$ 个样本对应聚类中心的标签，从而大幅度地减少了数据的存储内存，减少了运算量和运算量。

步骤 307，基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。

步骤 307 与前述实施例的步骤 205 基本一致，具体实现方式可以参考前述对步骤 205 的描述，此处不再赘述。

从图 3 中可以看出，与图 2 对应的实施例相比，本实施例中的构建检索数据库的方法突出了得到二级聚类中心以及得到第一数据集对应的码本的过程，从而对数据集进行了更精细的空间分割，减少了检索时延，也提升了检索的召回率，节省了内存。

继续参考图 4，图 4 示出了图 3 所示的构建检索数据库的方法的训练步骤的分解流程 400。该训练步骤分解如下：

步骤 401，计算第二数据与一级聚类中心和二级聚类中心的距离。

在本实施例中，针对所述第二数据集中的每个第二数据，计算每个第二

数据与各一级聚类中心以及各二级聚类中心的距离，其中，距离可为欧氏距离（L2 距离）或余弦距离（COS 距离），欧氏距离或余弦距离的计算方式为现有技术，在此不再赘述。

在本实施例的一些可选实施方式中，步骤 401 包括：计算第二数据与一级聚类中心的第一距离；响应于第一距离以及一级聚类中心与二级聚类中心的第二距离之间不满足三角不等式，进行剪枝操作。

在此实施例中，对于 L2/COS 距离空间下，任意三点满足三角不等式，在计算了第二数据与一级聚类中心的第一距离后，由于一级聚类中心与二级聚类中心之间的距离为已知，所以可根据三角不等式来进行剪枝操作，也即根据三角不等式，可预判第二数据与每个子空间的上界，按一定规则排序所有子空间，根据距离上界进行剪枝，从而大幅度降低整体计算量。三角不等式可表示为如下形式：

$$d(Q_x, Q_y) \leq d(Q_x, Q_z) + d(Q_z, Q_y);$$

其中， Q_x 、 Q_y 、 Q_z 代表空间中的任意三点，即上文中的第二数据、一级聚类中心以及二级聚类中心， d 代表任意两点之间的距离。

在一千万 SIFT1B 数据集上验证，该剪枝可剪掉 94% 的向量计算，极大地缩短了建库时间。此外，该剪枝方式可复用在训练和检索过程中任何需要计算样本与子空间距离的过程中。

在本实施例的一些可选实施方式中，步骤 401 还包括：响应于第一距离以及一级聚类中心与二级聚类中心的第二距离之间满足三角不等式，计算第二数据与二级聚类中心的距离。也即在第一距离以及一级聚类中心与二级聚类中心的第二距离之间满足三角不等式的情况下，才计算第二数据与二级聚类中心的距离。通过判断第一距离与第二距离是否满足三角不等式，从而减少了计算量。

步骤 402，基于计算结果确定与第二数据最接近的一级聚类中心和二级聚类中心。

在本实施例中，基于步骤 401 的计算结果确定与第二数据最接近的一级聚类中心和二级聚类中心。步骤 401 的计算结果越小，表明距离越近，从而确定最接近的一级聚类中心以及二级聚类中心。

步骤 403，针对第二数据集中的每个第二数据，基于与第二数据最接近的一级聚类中心和二级聚类中心的残差，得到各第二数据对应的残差向量。

在本实施例中，针对第二数据集中的每个第二数据，在确定了与第二数据最接近的一级聚类中心和二级聚类中心之后，可计算第二数据与其最接近的一级聚类中心和二级聚类中心的残差，从而得到各第二数据对应的残差向量。

步骤 404，将残差向量分为第一预设数目个子空间。

在本实施例中，可将步骤 403 得到的残差向量分为第一预设数目个子空间，其中，第一预设数目可根据实际情况进行设定。

步骤 405，针对第一预设数目个子空间中的每个子空间，计算子空间内的残差向量与子空间内的空间聚类中心的距离。

在本实施例中，针对第一预设数目个子空间中的每个子空间，可计算每个子空间内的残差向量与子空间内的空间聚类中心的距离，从而确定与该子空间内的残差向量距离最近的空间聚类中心，并记录该聚类中心的 ID。

步骤 406，用与子空间内的残差向量最接近的空间聚类中心的 ID 标记第二数据集中的每个第二数据，得到数据集对应的码本。

在本实施例中，可用步骤 405 确定的与子空间内的残差向量最接近的空间聚类中心的 ID 标记第二数据集中的每个第二数据，从而得到数据集对应的码本。

从图 4 中可以看出，该训练方法可以使用第一数据集中的各聚类中心的标识来标记第二数据集中的第二数据，从而使得在检索过程中，无需加载原始特征向量，极大地降低了检索过程所需的内存。

继续参考图 5，其示出了根据本公开的检索方法的一个实施例的流程 500。该检索方法包括以下步骤：

步骤 501，获取待检索数据。

在本实施例中，检索方法的执行主体（例如图 1 所示的服务器 105）可以获取待检索数据。其中，待检索数据可以为图像、视频和/或音频。可选地，上述执行主体在获取到待检索数据后，会提取待检索数据的特征，并对提取的特征进行归一化处理，以便更好地进行检索。

步骤 502，计算待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心。

在本实施例中，上述执行主体可以计算待检索数据与检索数据库中的各

个一级聚类中心的距离，并基于计算结果进行排序，例如进行升序排序，从而得到满足预设条件的第三预设数目个一级聚类中心，其中，第三预设数目可根据实际情况进行设定，例如，将第三预设数目设定为 10，也即根据升序排序结果，选取距离最近的 10 个一级聚类中心，未被选取的子空间在后续检索流程中则不会被访问，从而减少计算量，减少耗时。其中，检索数据库通过如前述实施例中的方法得到。

步骤 503，计算待检索数据与第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心。

在本实施例中，在确定了与待检索数据距离最近的第三预设数目个一级聚类中心后，上述执行主体可以计算待检索数据与第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，并对计算结果进行排序，例如进行升序排序，从而得到满足预设条件的第四预设数目个二级聚类中心，例如，将第四预设数目设定为 20，也即根据升序排序结果，选取距离最近的 20 个二级聚类中心，那么当一级聚类中心的个数为 10 时，此时一共得到了 $10 \times 20 = 200$ 个二级聚类中心。

步骤 504，基于检索数据库中的数据集中的码本，依次获取第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据。

在本实施例中，上述执行主体可以基于构建检索数据库时得到的数据集对应的码本，以优先队列的方式，从最近的二级聚类中心开始，依次取出其对应的样本数据。

步骤 505，计算待检索数据与样本数据的距离，基于计算结果返回第五预设数目个样本数据。

在本实施例中，上述执行主体可以计算待检索数据与步骤 504 取出的样本数据的距离，并将计算结果进行排序，从而返回第五预设数目个样本数据。第五预设数目可根据实际情况进行设定，假设将第五预设数目设置为 200，那么上述执行主体会基于排序结果，返回与待检索数据距离最近的前 200 个样本数据。

本公开实施例提供的检索方法，首先获取待检索数据；然后计算待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心；之后计算待检索数据与第三预设数目个一级聚类中心

中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；并基于检索数据库中的数据集中的码本，依次获取第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；最后计算待检索数据与样本数据的距离，基于计算结果返回第五预设数目个样本数据。本实施例中的检索方法，能够基于预先构建好的检索数据库进行检索，提高了检索召回率，并减少了检索时延，从而提高了检索效率。

进一步参考图 6，作为对上述各图所示方法的实现，本公开提供了一种构建检索数据库的装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 6 所示，本实施例的构建检索数据库的装置 600 包括：第一获取模块 601、第一聚类模块 602、第二聚类模块 603、得到模块 604 和训练模块 605。其中，第一获取模块 601，被配置成获取数据集，将数据集分为第一数据集和第二数据集，其中，第一数据集中的数据量小于第二数据集中的数据量；第一聚类模块 602，被配置成对第一数据集中的数据进行聚类，得到至少一个一级聚类中心；第二聚类模块 603，被配置成基于一级聚类中心对第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；得到模块 604，被配置成基于第一数据集中的数据与一级聚类中心以及二级聚类中心的残差，得到第一数据集对应的码本；训练模块 605，被配置成基于第一数据集对应的码本对第二数据集进行训练，得到数据集对应的码本。

在本实施例中，构建检索数据库的装置 600 中：第一获取模块 601、第一聚类模块 602、第二聚类模块 603、得到模块 604 和训练模块 605 的具体处理及其所带来的技术效果可分别参考图 2 对应实施例中的步骤 201-205 的相关说明，在此不再赘述。

在本实施例的一些可选的实现方式中，第二聚类模块包括：第一计算子模块，被配置成针对第一数据集中的每个第一数据，计算与第一数据最接近的一级聚类中心的残差，得到第一数据集对应的第一残差向量空间；聚类子模块，被配置成对第一残差向量空间进行聚类，得到对应的至少一个二级聚类中心。

在本实施例的一些可选的实现方式中，得到模块包括：第一得到子模块，被配置成针对第一数据集中的每个第一数据，基于与第一数据最接近的一级

聚类中心和二级聚类中心的残差,得到第一数据集对应的第二残差向量空间;量子子模块,被配置成对第二残差向量空间进行量化,得到第一数据集对应的码本。

在本实施例的一些可选的实现方式中,量子子模块包括:划分单元,被配置成将第二残差向量空间分为第一预设数目个子空间;聚类单元,被配置成分别对第一预设数目个子空间中的每个子空间进行聚类,得到第二预设数目个空间聚类中心;标记单元,被配置成分别对空间聚类中心的标识号 ID 进行标记,得到每个子空间的码本;直积单元,被配置成将每个子空间的码本做直积,得到第一数据集对应的码本。

在本实施例的一些可选的实现方式中,训练模块包括:第二得到子模块,被配置成针对第二数据集中的每个第二数据,基于与第二数据最接近的一级聚类中心和二级聚类中心的残差,得到各第二数据对应的残差向量;划分子模块,被配置成将残差向量分为第一预设数目个子空间;第二计算子模块,被配置成针对第一预设数目个子空间中的每个子空间,计算子空间内的残差向量与子空间内的空间聚类中心的距离;第三得到子模块,被配置成用与子空间内的残差向量最接近的空间聚类中心的 ID 标记第二数据集中的每个第二数据,得到数据集对应的码本。

在本实施例的一些可选的实现方式中,训练模块还包括:第三计算子模块,被配置成计算第二数据与一级聚类中心和二级聚类中心的距离;确定子模块,被配置成基于第三计算子模块的计算结果确定与第二数据最接近的一级聚类中心和二级聚类中心。

在本实施例的一些可选的实现方式中,第三计算子模块包括:第一计算单元,被配置成计算第二数据与一级聚类中心的第一距离;剪枝单元,被配置成响应于第一距离以及一级聚类中心与二级聚类中心的第二距离之间不满足三角不等式,进行剪枝操作。

在本实施例的一些可选的实现方式中,第三计算子模块还包括:第二计算单元,被配置成响应于第一距离以及一级聚类中心与二级聚类中心的第二距离之间满足三角不等式,计算第二数据与二级聚类中心的距离。

进一步参考图 7,作为对上述各图所示方法的实现,本公开提供了一种检索装置的一个实施例,该装置实施例与图 5 所示的方法实施例相对应,该

装置具体可以应用于各种电子设备中。

如图 7 所示，本实施例的检索装置 700 包括：第二获取模块 701、第一计算模块 702、第二计算模块 703、第三获取模块 704 和第三计算模块 705。其中，第二获取模块 701，被配置成获取待检索数据；第一计算模块 702，被配置成计算待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心；第二计算模块 703，被配置成计算待检索数据与第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；第三获取模块 704，被配置成基于检索数据库中的数据集中的数据集对应的码本，依次获取第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；第三计算模块 705，被配置成计算待检索数据与样本数据的距离，基于计算结果返回第五预设数目个样本数据。

在本实施例中，检索装置 700 中：第二获取模块 701、第一计算模块 702、第二计算模块 703、第三获取模块 704 和第三计算模块 705 的具体处理及其所带来的技术效果可分别参考图 5 对应实施例中的步骤 501-505 的相关说明，在此不再赘述。

根据本公开的实施例，本公开还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

图 8 示出了可以用来实施本公开的实施例的示例电子设备 800 的示意性框图。电子设备旨在表示各种形式的数字计算机，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

如图 8 所示，设备 800 包括计算单元 801，其可以根据存储在只读存储器（ROM）802 中的计算机程序或者从存储单元 808 加载到随机访问存储器（RAM）803 中的计算机程序，来执行各种适当的动作和处理。在 RAM 803 中，还可存储设备 800 操作所需的各种程序和数据。计算单元 801、ROM 802 以及 RAM 803 通过总线 804 彼此相连。输入/输出（I/O）接口 805 也连接至

总线 804。

设备 800 中的多个部件连接至 I/O 接口 805，包括：输入单元 806，例如键盘、鼠标等；输出单元 807，例如各种类型的显示器、扬声器等；存储单元 808，例如磁盘、光盘等；以及通信单元 809，例如网卡、调制解调器、无线通信收发机等。通信单元 809 允许设备 800 通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据。

计算单元 801 可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元 801 的一些示例包括但不限于中央处理单元（CPU）、图形处理单元（GPU）、各种专用的人工智能（AI）计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器（DSP）、以及任何适当的处理器、控制器、微控制器等。计算单元 801 执行上文所描述的各个方法和处理，例如构建检索数据库的方法或检索方法。例如，在一些实施例中，构建检索数据库的方法或检索方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元 808。在一些实施例中，计算机程序的部分或者全部可以经由 ROM 802 和/或通信单元 809 而被载入和/或安装到设备 800 上。当计算机程序加载到 RAM 803 并由计算单元 801 执行时，可以执行上文描述的构建检索数据库的方法或检索方法的一个或多个步骤。备选地，在其他实施例中，计算单元 801 可以通过其他任何适当的方式（例如，借助于固件）而被配置为执行构建检索数据库的方法或检索方法。

本文中以上描述的系统和技术和各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列（FPGA）、专用集成电路（ASIC）、专用标准产品（ASSP）、芯片上系统的系统（SOC）、负载可编程逻辑设备（CPLD）、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行

时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦除可编程只读存储器（EPROM 或快闪存储器）、光纤、便捷式紧凑盘只读存储器（CD-ROM）、光学储存设备、磁储存设备、或上述内容的任何合适组合。

为了提供与用户的交互，可以在计算机上实施此处描述的系统和技术，该计算机具有：用于向用户显示信息的显示装置（例如，CRT（阴极射线管）或者 LCD（液晶显示器）监视器）；以及键盘和指向装置（例如，鼠标或者轨迹球），用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的传感反馈（例如，视觉反馈、听觉反馈、或者触觉反馈）；并且可以用任何形式（包括声输入、语音输入或者、触觉输入）来接收来自用户的输入。

可以将此处描述的系统和技术实施在包括后台部件的计算系统（例如，作为数据服务器）、或者包括中间件部件的计算系统（例如，应用服务器）、或者包括前端部件的计算系统（例如，具有图形用户界面或者网络浏览器的用户计算机，用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互）、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信（例如，通信网络）来将系统的部件相互连接。通信网络的示例包括：局域网（LAN）、广域网（WAN）和互联网。

计算机系统可以包括客户端和服务端。客户端和服务端一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务端的关系。服务器可以是

云服务器，也可以为分布式系统的服务器，或者是结合了区块链的服务器。

应该理解，可以使用上面所示的各种形式的流程，重新排序、增加或删除步骤。例如，本发公开中记载的各步骤可以并行地执行也可以顺序地执行也可以不同的次序执行，只要能够实现本公开公开的技术方案所期望的结果，本文在此不进行限制。

上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，根据设计要求和因素，可以进行各种修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

权利要求书

1、一种构建检索数据库的方法，包括：

获取数据集，将所述数据集分为第一数据集和第二数据集，其中，所述第一数据集中的数据量小于所述第二数据集中的数据量；

对所述第一数据集中的数据进行聚类，得到至少一个一级聚类中心；

基于所述一级聚类中心对所述第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；

基于所述第一数据集中的数据与所述一级聚类中心以及所述二级聚类中心的残差，得到所述第一数据集对应的码本；

基于所述第一数据集对应的码本对所述第二数据集进行训练，得到所述数据集对应的码本。

2、根据权利要求1所述的方法，其中，所述基于所述一级聚类中心对所述第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心，包括：

针对所述第一数据集中的每个第一数据，计算与所述第一数据最接近的一级聚类中心的残差，得到所述第一数据集对应的第一残差向量空间；

对所述第一残差向量空间进行聚类，得到对应的至少一个二级聚类中心。

3、根据权利要求1或2所述的方法，其中，所述基于所述第一数据集中的数据与所述一级聚类中心以及所述二级聚类中心的残差，得到所述第一数据集对应的码本，包括：

针对所述第一数据集中的每个第一数据，基于与所述第一数据最接近的一级聚类中心和二级聚类中心的残差，得到所述第一数据集对应的第二残差向量空间；

对所述第二残差向量空间进行量化，得到所述第一数据集对应的码本。

4、根据权利要求3所述的方法，其中，所述对所述第二残差向量空间进行量化，得到所述第一数据集对应的码本，包括：

将所述第二残差向量空间分为第一预设数目个子空间；

分别对所述第一预设数目个子空间中的每个子空间进行聚类，得到第二预设数目个空间聚类中心；

分别对所述空间聚类中心的标识号 ID 进行标记，得到每个子空间的码本；

将所述每个子空间的码本做直积，得到所述第一数据集对应的码本。

5、根据权利要求 4 所述的方法，其中，所述基于所述第一数据集对应的码本对所述第二数据集进行训练，得到所述数据集对应的码本，包括：

针对所述第二数据集中的每个第二数据，基于与所述第二数据最接近的一级聚类中心和二级聚类中心的残差，得到各第二数据对应的残差向量；

将所述残差向量分为第一预设数目个子空间；

针对所述第一预设数目个子空间中的每个子空间，计算所述子空间内的残差向量与所述子空间内的空间聚类中心的距离；

用与所述子空间内的残差向量最接近的空间聚类中心的 ID 标记所述第二数据集中的每个第二数据，得到所述数据集对应的码本。

6、根据权利要求 5 所述的方法，其中，在所述基于与所述第二数据最接近的一级聚类中心和二级聚类中心的残差，得到各第二数据对应的残差向量之前，所述方法还包括：

计算所述第二数据与所述一级聚类中心和所述二级聚类中心的距离；

基于计算结果确定与所述第二数据最接近的一级聚类中心和二级聚类中心。

7、根据权利要求 6 所述的方法，其中，所述计算所述第二数据与所述一级聚类中心和所述二级聚类中心的距离，包括：

计算所述第二数据与所述一级聚类中心的第一距离；

响应于所述第一距离以及所述一级聚类中心与所述二级聚类中心的第二距离之间不满足三角不等式，进行剪枝操作。

8、根据权利要求 7 所述的方法，其中，所述计算所述第二数据与所述一级聚类中心和所述二级聚类中心的距离，还包括：

响应于所述第一距离以及所述一级聚类中心与所述二级聚类中心的第二距离之间满足三角不等式，计算所述第二数据与所述二级聚类中心的距离。

9、一种检索方法，包括：

获取待检索数据；

计算所述待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心，其中，所述检索数据库通过权利要求 1-8 中任一项所述的方法得到；

计算所述待检索数据与所述第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；

基于所述检索数据库中的所述数据集对应的码本，依次获取所述第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；

计算所述待检索数据与所述样本数据的距离，基于计算结果返回第五预设数目个样本数据。

10、一种构建检索数据库的装置，包括：

第一获取模块，被配置成获取数据集，将所述数据集分为第一数据集和第二数据集，其中，所述第一数据集中的数据量小于所述第二数据集中的数据量；

第一聚类模块，被配置成对所述第一数据集中的数据进行聚类，得到至少一个一级聚类中心；

第二聚类模块，被配置成基于所述一级聚类中心对所述第一数据集中的数据进行聚类，得到对应的至少一个二级聚类中心；

得到模块，被配置成基于所述第一数据集中的数据与所述一级聚类中心以及所述二级聚类中心的残差，得到所述第一数据集对应的码本；

训练模块，被配置成基于所述第一数据集对应的码本对所述第二数据集进行训练，得到所述数据集对应的码本。

11、根据权利要求 10 所述的装置，其中，所述第二聚类模块包括：

第一计算子模块，被配置成针对所述第一数据集中的每个第一数据，计算与所述第一数据最接近的一级聚类中心的残差，得到所述第一数据集对应的第一残差向量空间；

聚类子模块，被配置成对所述第一残差向量空间进行聚类，得到对应的至少一个二级聚类中心。

12、根据权利要求 10 或 11 所述的装置，其中，所述得到模块包括：

第一得到子模块，被配置成针对所述第一数据集中的每个第一数据，基于与所述第一数据最接近的一级聚类中心和二级聚类中心的残差，得到所述第一数据集对应的第二残差向量空间；

量化子模块，被配置成对所述第二残差向量空间进行量化，得到所述第一数据集对应的码本。

13、根据权利要求 12 所述的装置，其中，所述量化子模块包括：

划分单元，被配置成将所述第二残差向量空间分为第一预设数目个子空间；

聚类单元，被配置成分别对所述第一预设数目个子空间中的每个子空间进行聚类，得到第二预设数目个空间聚类中心；

标记单元，被配置成分别对所述空间聚类中心的标识号 ID 进行标记，得到每个子空间的码本；

直积单元，被配置成将所述每个子空间的码本做直积，得到所述第一数据集对应的码本。

14、根据权利要求 13 所述的装置，其中，所述训练模块包括：

第二得到子模块，被配置成针对所述第二数据集中的每个第二数据，基于与所述第二数据最接近的一级聚类中心和二级聚类中心的残差，得到各第二数据对应的残差向量；

划分子模块，被配置成将所述残差向量分为第一预设数目个子空间；

第二计算子模块，被配置成针对所述第一预设数目个子空间中的每个子空间，计算所述子空间内的残差向量与所述子空间内的空间聚类中心的距离；

第三得到子模块，被配置成用与所述子空间内的残差向量最接近的空间聚类中心的 ID 标记所述第二数据集中的每个第二数据，得到所述数据集对应的码本。

15、根据权利要求 14 所述的装置，其中，所述训练模块还包括：

第三计算子模块，被配置成计算所述第二数据与所述一级聚类中心和所述二级聚类中心的距离；

确定子模块，被配置成基于所述第三计算子模块的计算结果确定与所述第二数据最接近的一级聚类中心和二级聚类中心。

16、根据权利要求 15 所述的装置，其中，所述第三计算子模块包括：

第一计算单元，被配置成计算所述第二数据与所述一级聚类中心的第一距离；

剪枝单元，被配置成响应于所述第一距离以及所述一级聚类中心与所述二级聚类中心的第二距离之间不满足三角不等式，进行剪枝操作。

17、根据权利要求 16 所述的装置，其中，所述第三计算子模块还包括：

第二计算单元，被配置成响应于所述第一距离以及所述一级聚类中心与所述二级聚类中心的第二距离之间满足三角不等式，计算所述第二数据与所述二级聚类中心的距离。

18、一种检索装置，包括：

第二获取模块，被配置成获取待检索数据；

第一计算模块，被配置成计算所述待检索数据与检索数据库中的一级聚类中心的距离，得到满足预设条件的第三预设数目个一级聚类中心，其中，所述检索数据库通过权利要求 1-8 中任一项所述的方法得到；

第二计算模块，被配置成计算所述待检索数据与所述第三预设数目个一级聚类中心中的每个一级聚类中心下的二级聚类中心的距离，得到满足预设条件的第四预设数目个二级聚类中心；

第三获取模块，被配置成基于所述检索数据库中的数据集中的码本，依次获取所述第四预设数目个二级聚类中心中的每个二级聚类中心对应的样本数据；

第三计算模块，被配置成计算所述待检索数据与所述样本数据的距离，基于计算结果返回第五预设数目个样本数据。

19、一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中，

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求 1-9 中任一项所述的方法。

20、一种存储有计算机指令的非瞬时计算机可读存储介质，所述计算机指令用于使所述计算机执行权利要求 1-9 中任一项所述的方法。

21、一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求 1-9 中任一项所述的方法。

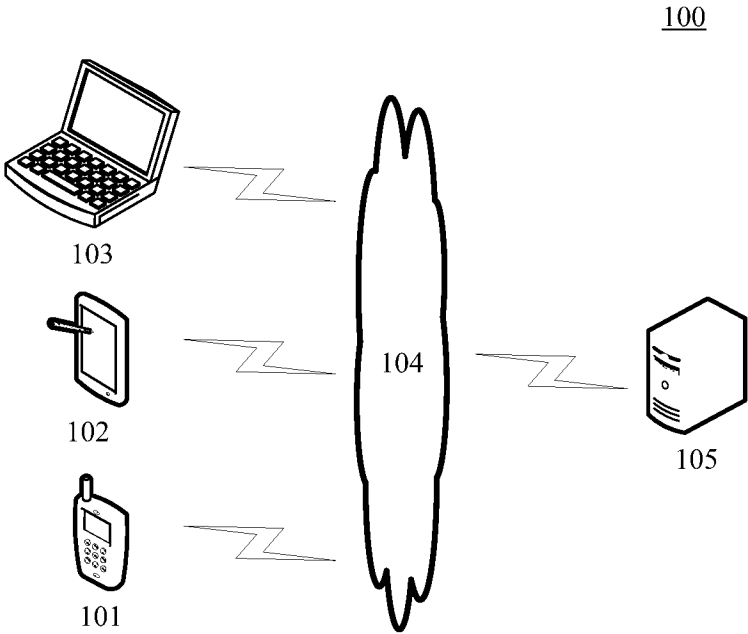


图 1

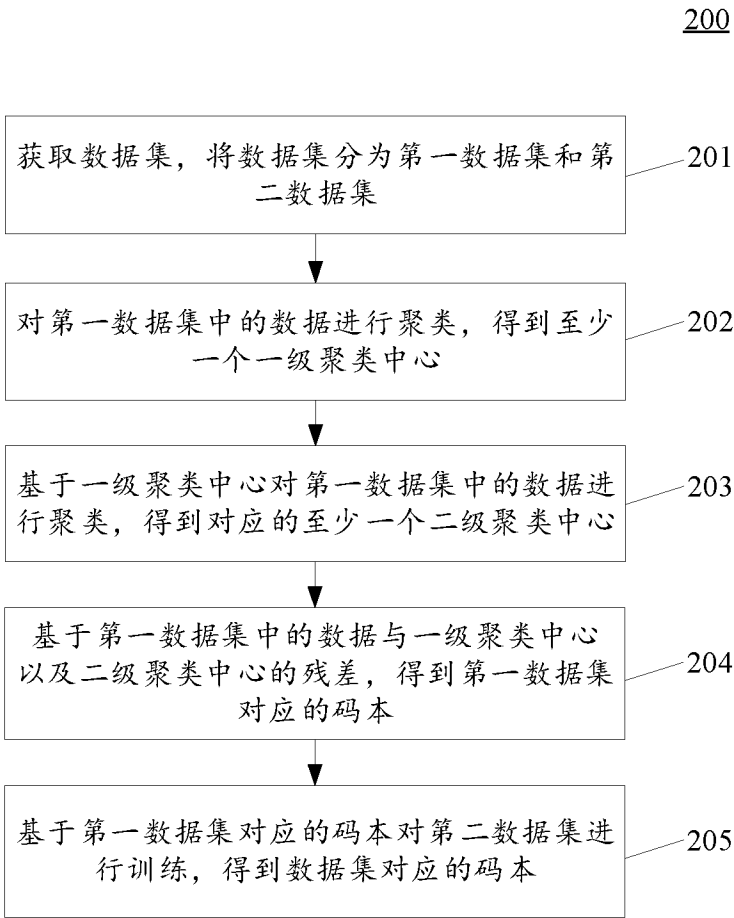


图 2

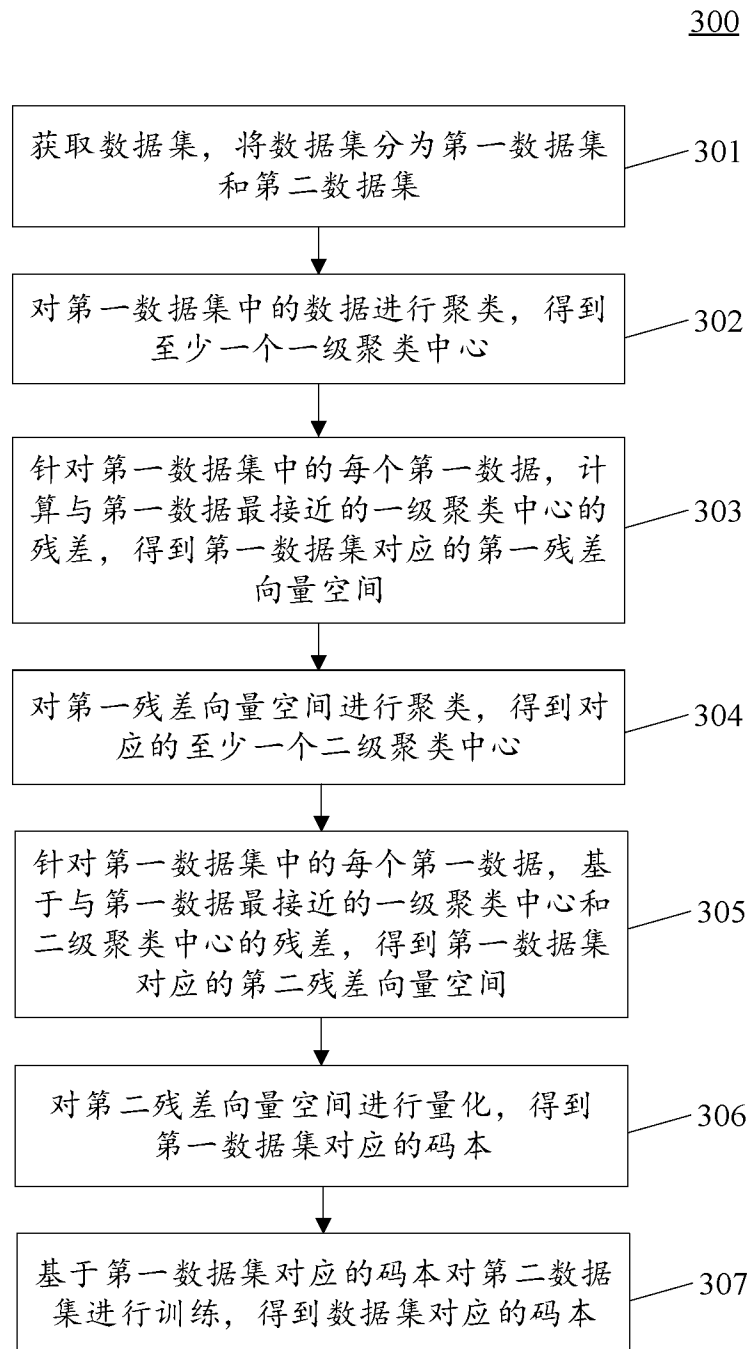


图 3

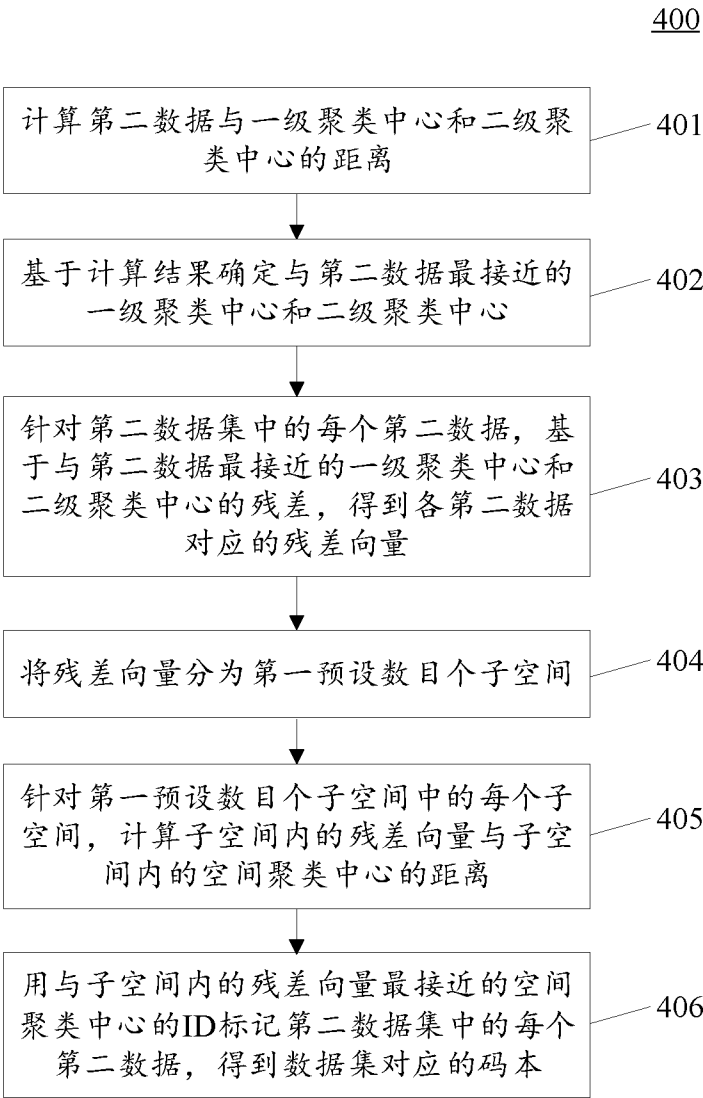


图 4

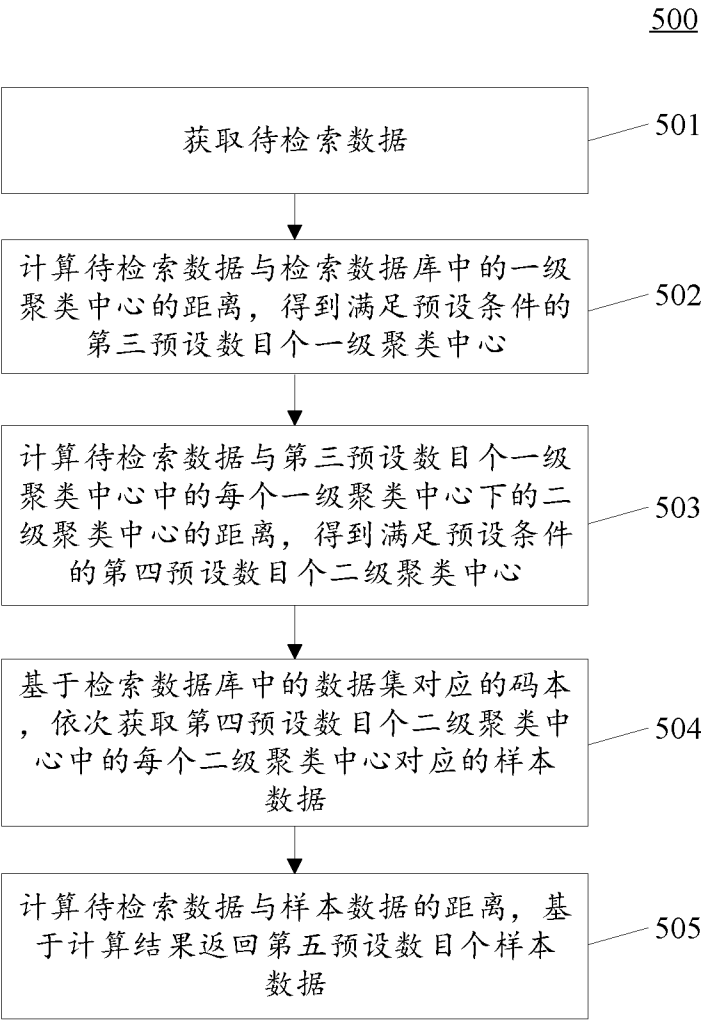


图 5

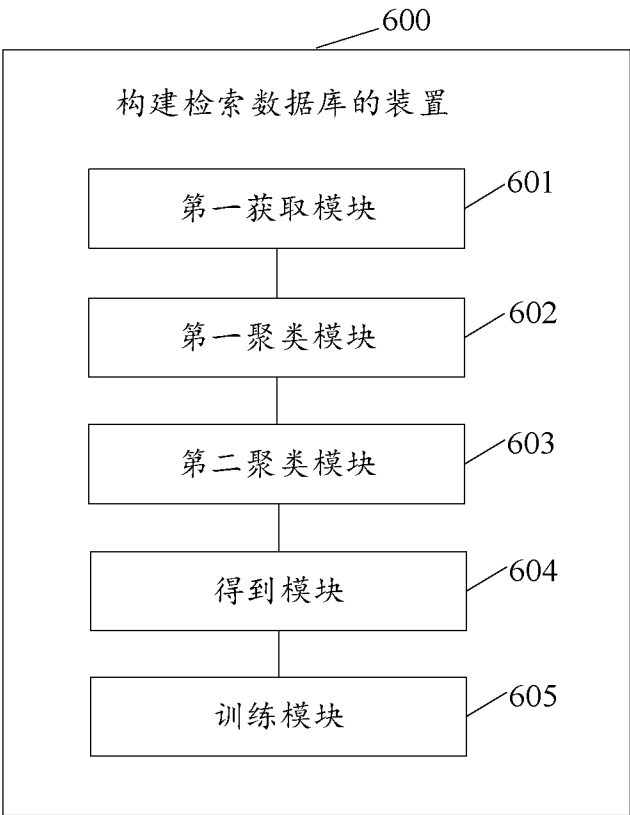


图 6

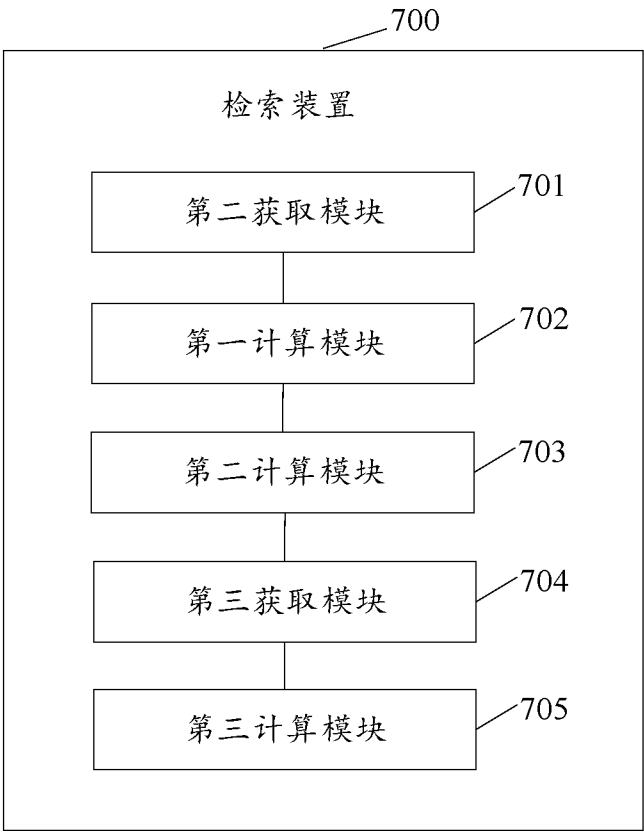


图 7

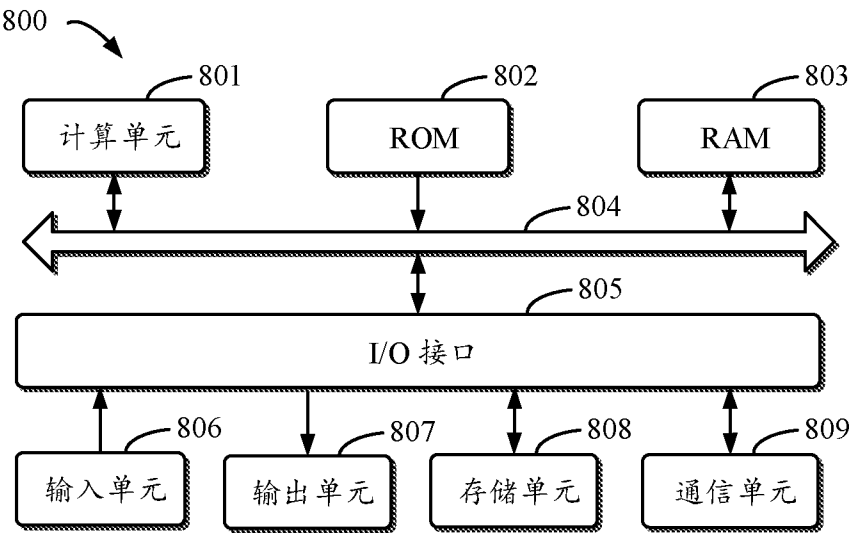


图 8

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2022/081154

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/21(2019.01)i; G06K 9/62(2022.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F; G06K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNTXT, CNABS, ENTXT, ENTXTC, DWPI, CNKI: 检索, 数据集, 聚类, 第一, 第二, 残差, 码本, 训练, 分层, 层次, 向量, search, retrieve, data set, cluster, first, second, residual, code book, train, hierarchy, vector

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 113656373 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.) 16 November 2021 (2021-11-16) claims 1-21	1-21
X	CN 110168525 A (GOOGLE INC.) 23 August 2019 (2019-08-23) description, paragraphs 28-102	1-21
X	CN 110825894 A (PING AN TECHNOLOGY (SHENZHEN) CO., LTD.) 21 February 2020 (2020-02-21) description, paragraphs [0014]-[0029]	1-21
A	CN 110134804 A (BEIJING DAJIA INTERCONNECTION INFORMATION TECHNOLOGY CO., LTD.) 16 August 2019 (2019-08-16) entire document	1-21
A	CN 111444363 A (GOSUNCN TECHNOLOGY GROUP CO., LTD.) 24 July 2020 (2020-07-24) entire document	1-21
A	CN 110609916 A (SICHUAN DONGFANG WANGLI TECHNOLOGY CO., LTD.) 24 December 2019 (2019-12-24) entire document	1-21



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

25 May 2022

Date of mailing of the international search report

02 June 2022

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/
CN)
No. 6, Xitucheng Road, Jimenqiao, Haidian District, Beijing
100088, China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2022/081154

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	113656373	A	16 November 2021	None			
CN	110168525	A	23 August 2019	US	2018101570	A1	12 April 2018
				EP	3526689	A1	21 August 2019
				WO	2018071148	A1	19 April 2018
				US	10719509	B2	21 July 2020
CN	110825894	A	21 February 2020	None			
CN	110134804	A	16 August 2019	CN	110134804	B	17 September 2021
CN	111444363	A	24 July 2020	None			
CN	110609916	A	24 December 2019	None			

国际检索报告

国际申请号

PCT/CN2022/081154

A. 主题的分类

G06F 16/21 (2019.01) i; G06K 9/62 (2022.01) i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06F; G06K

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用))

CNTXT, CNABS, ENTXT, ENTXTC, DWPI, CNKI: 检索, 数据集, 聚类, 第一, 第二, 残差, 码本, 训练, 分层, 层次, 向量, search, retrieve, data set, cluster, first, second, residual, code book, train, hierarchy, vector

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
PX	CN 113656373 A (百度在线网络技术北京有限公司) 2021年11月16日 (2021 - 11 - 16) 权利要求1-21	1-21
X	CN 110168525 A (谷歌有限责任公司) 2019年8月23日 (2019 - 08 - 23) 说明书第28-102段	1-21
X	CN 110825894 A (平安科技深圳有限公司) 2020年2月21日 (2020 - 02 - 21) 说明书第14-29段	1-21
A	CN 110134804 A (北京达佳互联信息技术有限公司) 2019年8月16日 (2019 - 08 - 16) 全文	1-21
A	CN 111444363 A (高新兴科技集团股份有限公司) 2020年7月24日 (2020 - 07 - 24) 全文	1-21
A	CN 110609916 A (四川东方网力科技有限公司) 2019年12月24日 (2019 - 12 - 24) 全文	1-21

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2022年5月25日

国际检索报告邮寄日期

2022年6月2日

ISA/CN的名称和邮寄地址

中国国家知识产权局(ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

王雪莲

电话号码 62411747

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2022/081154

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	113656373	A	2021年11月16日	无			
CN	110168525	A	2019年8月23日	US	2018101570	A1	2018年4月12日
				EP	3526689	A1	2019年8月21日
				WO	2018071148	A1	2018年4月19日
				US	10719509	B2	2020年7月21日
CN	110825894	A	2020年2月21日	无			
CN	110134804	A	2019年8月16日	CN	110134804	B	2021年9月17日
CN	111444363	A	2020年7月24日	无			
CN	110609916	A	2019年12月24日	无			