# **ORCV 向量文件格式定义**

ORCV：Orcadt Computer Vision

1、介绍

和类似系统的demo版本只有一个索引文件的设计不同，ORCV索引采用了三个数据文件的设计。

* header信息，头部，记录向量数和其它各种选项，这个文件在运行过程中是变化的
* 每个centriod的归一化值，这是静态的，因为centriod在k-means算法跑完给定的数据后就确定的。系统运行过程中，只是使用k-means算法得到的centriod，并不更新。
* 向量，每个向量一条记录（添加向量的操作会向这个文件尾部追加记录）

header和centriod归一化值是在训练过程中创建的，初始版本n = 0，向量文件长度为0，被视作基础索引。

1.1 索引头部定义

索引头部通过结构体orcvhdr\_t定义，结构如下：

**typedef struct orcvhdr {**

**uint32\_t n; // number of vectors (changes)**

**uint32\_t nc; // number of centroids (fixed)**

**uint32\_t code\_size; // code size in PQ format**

**uint32\_t code\_bytes; // code size in bytes**

**uint32\_t d; // vector dimensions**

**uint32\_t M; // seach index internal**

**uint32\_t efConstruction; // search index internal**

**float32\_t dmatch; // distance for vector match**

**float32\_t dnear; // distance for vector near match**

**uint8\_t do\_opq; // if vector rotation matrix is used**

**} orcvhdr\_t;**

在训练结束，开始启动索引服务前，必须创建好这个文件。

这些文件通过ivf-hnsw/examples/run\_\*.sh 脚本创建。

orcvhdr\_t 结构体的成员参考自 Parser.h。

1.2 归一化的centriod文件

理解centriod文件的格式，可以阅读代码 index->write()

这个文件格式比较简单，开始是4字节长度（centriod的数目），后面是和长度信息匹配的centriod数据，每个centriod数据是一个float类型的浮点数。

1.3 向量文件格式

每个向量包含：

8 bytes的EID，这是为向量分配的全局应用ID

4 bytes的centriod ID

PQ字节形式的向量（PQ长度为16时对应2字节）

PQ字节形式的向量归一化（1字节）

注意：这个文件没有内部向量id，这个文件是个数组，向量ID是数组索引，用于得到文件索引。

1.4 Home目录

索引服务的环境变量包括home目录和监听端口。在home目录下，有一系列文件：

**sprintf(env.idxpath, "%s/idx.vec", env.home);**

**sprintf(env.hdrpath, "%s/hdr.vec", env.home);**

**sprintf(env.centidpath, "%s/centid.vec", env.home);**

**sprintf(env.cnormspath, "%s/cnorms.vec", env.home);**

**sprintf(env.pqpath, "%s/pq.vec", env.home);**

**sprintf(env.opqpath, "%s/opq.vec", env.home);**

**sprintf(env.normpqpath, "%s/normpq.vec", env.home);**

**sprintf(env.centpath, "%s/cent.vec", env.home);**

**sprintf(env.edgepath, "%s/edge.vec", env.home);**

**sprintf(env.nodepath, "%s/node.vec", env.home);**

注意：我们希望文件扩展名和读写这类文件的工具完全匹配。就是说，根据文件扩展名，匹配文件格式，确定读取文件的方式。

通过上面定义的索引组件，别的没有提到的文件和ivf-hnsw中benchmarks涉及的文件格式完全一致，也是基础索引文件集合的一部分。ORCV只是标准化了路径命名，而不是使用命令行选项。