基于 Hadamard 变换的高维图像检索方法*)

崔江涛　周水生　周利华

(西安电子科技大学计算机网络与信息安全教育部重点实验室　西安 710071)

摘　要　传统索引方法对高维数据进行近邻搜索时会面临维数灾难问题，向量近似方法是一种有效的高维检索方法。提出一种 Hadamard 变换域上的向量近似方法，在变换域能量最大的分量上建立顺序索引，然后建立近似向量文件。同时提出低维过滤算法，可以在近邻搜索过程中高效排除不匹配近似向量，减少 I/O 访问时间，提高查询效率。在大型高维图像特征库上的实验表明，该方法性能优于小波变换域的向量近似方法。

关键词　图像数据库，维数灾难，k-近邻搜索，向量近似，Hadamard 变换

A New Indexing Method for High-Dimensional Image Databases Using Hadamard Transform

CUI Jiang-Tao　ZHOU Shui-Sheng　ZHOU Li-Hua

(Key Lab. of Computer Network and Information Security under the Ministry of Education, Xidian University, Xi'an 710071)

Abstract　Traditional indexing methods face the difficulty of 'curse of dimensionality' at high dimensionality. The vector approximation file(VA-File) approach based on wavelet transform is a very efficient high-dimensional indexing method. In this paper, a new VA-File approach in the Hadamard transform domain is introduced. This approach combines Hadamard transform and principle component filtering algorithm, which can reduce the searching complexity and I/O cost dramatically on large image databases. Experiment results show that the new method is more efficient than VA-File based on wavelet transform.

Keywords　Image databases, Curse of dimensionality, k-nearest neighbor search, Vector approximation, Hadamard transform

基于内容的图像检索(content-based image retrieval，简称 CBIR)已经成为图像数据库中的一项重要应用。CBIR 系统通常采用一定维数的向量来表示图像特征，通过某种距离度量方式计算两个向量之间的距离可以衡量图像之间的相似性，所以图像的相似性查询就相当于向量空间中的 k 近邻搜索问题。用于表示图像特征的向量往往具有高维特性，高维索引机制也就成为最近几年多媒体领域的一个研究热点[1]。最近的研究表明，传统的多维检索技术(R^* 树，X 树，SR 树和网格文件等)[2]在高维情况下(超过几十维)检索效率很低，其性能甚至低于最简单的顺序查找算法[3]，这种现象又被称为维数灾难(curse of dimensionality)。

为解决上述问题，研究者提出许多针对高维数据的检索方法，包括维数缩减方法[4]，向量近似方法[3]和近似近邻搜索方法[5]等。向量近似方法(Vector Approximation file，简称 VA-file)由 Weber 等人在 1998 年提出，对于精确近邻搜索，它是目前唯一的在高维情况下优于顺序查找的一种检索结构，其基本思想是将高维特征数据进行压缩并近似存储，通过过滤数据点来加快搜索速度。最近研究者对 VA-File 加以改进，提出正交变换域中的 VA-File 方法。正交变换一般都具有能量集中特性，可以消除各维数据之间的相关性。文[6]中提出了基于 K-L 变换的 VA-File 方法，但是该方法不能实际应用，能力不强，首先对于大型的高维图像数据库，K-L 变换的运算复杂度太高；其次它不能支持数据点的动态插入和删除操作，当数据动态变动时，需要重新进行 K-L 变换计算。笔者在文[7]中提出了基于小波变换的 VA-File 方法，为计算简单，采用了 Haar 小波变换。在进一步研究的基础上，笔者采用 Hadamard 变换代替小波变换，建立 Hadamard 变换域上的 VA-File 方法。与小波变换相比，Hadamard 变换不但计算复杂度更低，而且具有更好的能力集中特性。同时，本文还在能量(方差)最大的分量上建立顺序索引，引入低维过滤算法，可以进一步降低 k 近邻搜索过程的查询时间和 I/O 时间。

1　相关工作介绍

传统的基于树型结构的多维检索方法在高维情况下性能会低于最简单的顺序查找算法，这就是维数灾难现象，而 I/O 复杂度则是检索过程中的性能瓶颈。在 VA-File 方法中，为减少 I/O 访问时间，对原特征向量进行近似(压缩)存储。在进行近邻搜索时，通过在近似向量上的距离运算可以过滤掉大多数数据点，搜索过程中只需访问少数原始特征向量。

在 d 维向量空间内对向量进行近似时，每一维空间 j 分配一定的近似位数 b_j，将第 j 维坐标轴分割成 2^{b_j} 个区间，$\sum_{j=1}^{d} b_j = b$。按上述分割方法，整个向量空间可以分隔成 2^b 个互不相连的超立方体。一个特征向量可由所处的超立方体近似表示，称为近似向量，其二进制长度为 b。图 1 中给出一个简单示例，对二维空间中每一维空间各分配 2 位近似位数，整个空间可以分为 16 个正方形，那么向量 p_i 的近似向量是

*)基金项目：(十五国防科技(电子)预研项目，413160501)。崔江涛　博士研究生，讲师，主要研究方向为图像处理，多媒体数据库，高维检索技术等。周水生　博士研究生，讲师，主要研究方向为优化算法，模式识别，图像处理等。周利华　教授，博士生导师，主要研究方向为多媒体技术，网络安全，高精度大幅面扫描技术。

10 11。所有的近似向量顺序排列组成的序列就是一个向量近似文件 VA-file。根据 p_i 的近似向量，可以计算向量 q 与 p_i 所在的立方体之间的距离上界 u_i 和下界 l_i，如图 1 所示。

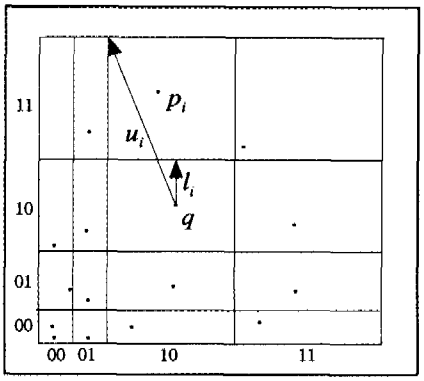

图 1　VA-File 空间分割示意图

k 近邻搜索包括两个阶段，第一阶段根据近似向量 a_i 计算特征向量 p_i 与查询向量 q 的距离下界 l_i 和距离上界 u_i，当 l_i 大于目前的第 k 小的距离上界时，就可以把此向量过滤掉，因为目前已经找到了至少 k 个更好的候选向量。没有过滤掉的近似向量作为候选向量进入第二阶段。第二阶段中需要访问原始特征向量，根据原始特征向量计算得到最终的 k 个近邻。

将第 j 维坐标轴的 2^{b_j} 个区间分割点标记为 $f_{l,j}$，$l=0,1,\cdots,2^{b_j}-1$。第 l 个分割区间可以标记为 $[f_{l,j}, f_{l+1,j}]$。对于 L_p 型距离测度，u_i 和 l_i 的计算公式如下：

$$l_i=\left(\sum_{j=1}^{d}(l_{i,j})^p\right)^{1/p}$$

$$u_i=\left(\sum_{j=1}^{d}(u_{i,j})^p\right)^{1/p} \tag{1}$$

其中 $l_{i,j}=\begin{cases} q_j-f_{l+1,j} & q_j>f_{l+1,j} \\ 0 & q_j\in[f_{l,j},f_{l+1,j}] \\ f_{l,j}-q_j & q_j<f_{l,j} \end{cases}$

$$u_{i,j}=\begin{cases} q_j-f_{l,j} & q_j>f_{l+1,j} \\ \max(q_j-f_{l,j},f_{l+1,j}-q_j) & q_j\in[f_{l,j},f_{l+1,j}] \\ f_{l+1,j}-q_j & q_j<f_{l,j} \end{cases}$$

针对 VA-File 方法的改进主要包括三个方面[6]：1)采用 Lloyd's 算法非均匀分割坐标轴，可以提高近似性能。2)在正交变换域中建立 VA-File，通过正交变换可以消除数据点之间的相关性。3)根据每一维空间能量(方差)不同，分配不同的量化位数。

2　基于 Hadamard 变换的检索结构

令 H_n 为 $2^n\times 2^n$ 的 Hadamard 变换矩阵，其元素属于集合{1，−1}。假定 $d=2^n$，有以下基本定义和性质：

$$H_1=\begin{Bmatrix}1 & 1\\ 1 & -1\end{Bmatrix},\ H_{n+1}=\begin{Bmatrix}H_n & H_n\\ H_n & -H_n\end{Bmatrix}$$

定理 1　$H_nH_n=dI_d$，其中 I_d 是 d 阶单位矩阵。

设 d 维向量 q 和 p_i，q 和 p_i 的 Hadamard 变换域矢量分别为 Q 和 P_i，即 $Q=H_nq$，$P_i=H_np_i$。

定理 2　$(\|Q\|_2)^2=d\cdot(\|q\|_2)^2$

证明：$(\|Q\|_2)^2=Q^TQ=(H_nq)^T(H_nq)=q^TH_nH_nq=q^TdI_dq=dq^Tq=d(\|q\|_2)^2$。证毕。

定理 3　$L_2(Q,P_i)=d\cdot L_2(q,p_i)$

证明：令 $w=q-p_i$，$W=H_nw$。则 $(\|w\|_2)^2=L_2(q,p_i)$，$(\|W\|_2)^2=L_2(Q,P_i)$。由定理 3 可得 $L_2(Q,P_i)=(\|W\|_2)^2=d\cdot(\|w\|_2)^2=d\cdot L_2(q,p_i)$。证毕。

由 Hadamard 变换域中的向量 P_i 生成近似向量，Q_l 和 $P_{i,1}$ 分别是向量 Q 和 P_i 的第一维分量，$l_{i,1}$ 和 $u_{i,l}$ 分别表示由向量 $P_{i,1}$ 和 Q_l 计算所得的距离下界和上界，可得以下定理。

定理 4　采用 L_2 度量距离时，$l_{i,1}\leqslant l_i\leqslant L_2(Q,P_i)$。

由 l_i 的定义，即 $l_i\leqslant L_p(Q,P_i)\leqslant u_i$ 和式(1)可以推导出上述定理。由定理 3 可知，Hadamard 虽然不是正交变换，但是变换前后的能量成倍数关系，因此在变换域中搜索最近邻与在空间(时)域中搜索是等价的，所以我们可以在 Hadamard 变换域中建立向量近似文件。

在离线计算近似向量的过程中，首先计算所有向量 p_i 的变换域向量 P_i，按照 P_i 第一个分量的大小 $P_{i,1}$ 对向量进行排序，然后对变换域向量 P_i 计算近似向量 a_i，近似向量的排列顺序与变换域向量一致。上述步骤相当于对第一维分量建立了顺序索引。

近似向量文件建立过程：

步骤 1：对原始特征向量 p_i 进行 Hadamard 变换，得到变换域向量 P_i。

步骤 2：按照 $P_{i,1}$ 值对向量 P_i 进行升序排列。

步骤 3：计算 P_i 各维分量的方差，按照文[7]中提供的算法对各个分量分配近似位数。

步骤 4：采用 Lloyd's 分割算法分割坐标轴，根据 P_i 计算近似向量 a_i。

3　近邻搜索算法

在 VA-File 方法的近邻搜索过程中，首先根据近似向量计算距离上下界，如果距离下界大于目前第 k 小的距离上界，那么可以将此向量过滤掉，因为已经至少有 k 个更好的候选向量。上述过程中，需要访问所有的近似向量。针对 Hadamard 变换域的近似向量建立过程，提出一种新的 k 近邻搜索算法，可以减少近似向量的访问数量，即降低近似向量的 I/O 访问时间，又提高搜索效率。

进行 k 近邻搜索搜索时，首先根据查询向量 q 的变换域向量 Q 计算近似向量 a_q，$a_{q,1}$ 是 a_q 的第一维分量。选择第一维分量与 $a_{q,1}$ 相同的近似向量作为初始匹配近似向量，即 $a_{i,1}=q_{q,1}$。由 Lloyd's 分割算法可知，坐标轴上各个分割区间内的数据点数基本相同，所以可以找到初始匹配近似向量，如果初始匹配的近似向量个数大于 1，选择序列中间位置的近似向量作为初始匹配向量。从初始匹配近似向量所处的数据页面开始，采用升序和降序交叉访问的次序访问数据页面，访问顺序如图 2 所示。

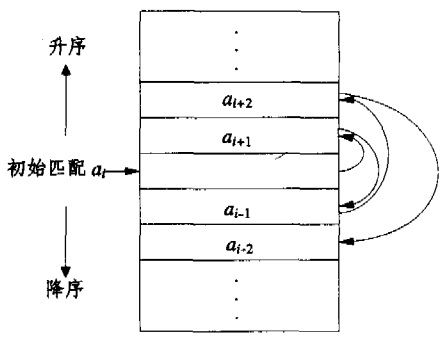

图 2　近邻搜索近似向量访问顺序

用数组 d_{max} 保存目前最小的 k 个距离上界，第 k 小的距离上界为 $d_{max}[k]$。以升序访问为例，假定在升序访问时下一个要匹配的近似向量为 a_m，计算距离下界 l_m 的第一个分量 $l_{m,1}$，如果 $l_{m,1}\geqslant d_{max}[k]$，则排除该向量，同时可以在升序访问顺序上终于访问，因为对任意 a_i，$m<i\leqslant N$，$l_i\geqslant l_{i,1}\geqslant d_{max}[k]$ 成立，可以排除向量 P_i。证明如下：

$l_{m,1} \geqslant d_{max}[k]$，$d_{max}[k]$是大于 0 的实数，显而易见，$Q_1 < P_{m,1}$成立。

由近似向量的排列顺序可知，$Q_1 < P_{m,1} \leqslant P_{i,1}$，$m < i \leqslant N$，即 $L_2(Q_1, P_{m,1}) \leqslant L_2(Q_1, P_{i,1})$，那么 $L_2(Q_1, P_{i,1}) \geqslant L_2(Q_1, P_{m,1}) \geqslant l_m \geqslant l_{m,1} \geqslant d_{max}[k]$成立，即 $l_{i,1} \geqslant d_{max}[k]$，证毕。

在计算过程中，如果 $l_{i,1} < d_{max}$，则继续计算 l_i，如果 $l_i \geqslant d_{max}$，则排除该向量。否则计算 u_i，并且用 u_i 更新 d_{max}。下面给出 k 近邻搜索过滤算法的详细描述，过滤结束后候选向量的计算过程与 VA-file 方法相同。

步骤 1：初始化数组 $d_{max}[k]$，数组中的距离设为 *MAXREAL*；升序访问近似向量时，用 a_i 表示，降序访问时用 a_j 表示。

步骤 2：计算查询向量 q 的变换域向量 Q，建立近似向量 a_q，查找初始匹配近似向量 a_m。

步骤 3：判定目前的访问顺序和需要访问的近似向量，以升序访问为例，假设要访问的近似向量是 a_i，根据 a_i 计算 $l_{i,1}$。如果 $l_{m,1} \geqslant d_{max}[k]$，则执行步骤(5)。否则，计算 l_i，如果 $l_i \geqslant d_{max}$，则执行步骤(4)。否则计算 u_i，更新数组 $d_{max}[k]$，向量 P_i 作为候选向量，进入步骤(4)。

步骤 4：如果 i 等于 N，则升序访问结束；如果 j 等于 1，降序访问结束；如果升序和降序访问都结束，则算法结束。否则，如果升序访问，则 i 值加 1，降序访问，j 值减 1，进入步骤(3)。

步骤 5：如果 $l_{i,1} \geqslant d_{max}$，则升序访问结束；如果 $l_{j,1} \geqslant d_{max}$，则降序访问结束。如果升序访问和降序访问都结束，算法结束，否则进入步骤(4)。

4 实验及算法性能评价

采用一个通用的高维特征集对文中涉及的检索方法进行性能测试。该数据集是一个航拍图像库，包含 275245 幅图像，使用 Gabor 纹理作为高维图像特征[8]，特征维数为 60，我们将其扩展为 64 维。采用 HP 工作站 XW4100(PIV3.0GHz CPU，1GB 主存)作为实验硬件平台，软件运行环境为微软的 Windows XP 操作系统。从数据集中随机选取 100 个特征向量作为 k 近邻搜索的查询向量，最后所得的测试结果是 100 个查询向量性能数据的平均值。

本文首先实现了基于 Haar 小波变换的 VA-File 方法(WTVA)和基于 Hadamard 变换的 VA-File 方法(HTVA)，为了比较小波变换和 Hadamard 变换在向量近似方法中的应用，这两种方法中近似向量都没有根据第一维分量进行排序。近似向量的平均位串长度实现了 2 位/维、4 位/维和 6 位/维这 3 种方案。近邻搜索分别实现了 10-NN、20-NN 和 50-NN 近邻搜索。表 1 给出了两种方法在近邻搜索过程中的第一阶段过滤效率(候选向量比例)。

表 1 WTVA 和 HTVA 的第一阶段过滤效率(百分比)

		2bits	4bits	6bits
10-NN	WTVA	90.54	5.63	0.38
	HTVA	90.28	5.21	0.37
20-NN	WTVA	90.57	7.36	0.53
	HTVA	90.31	6.04	0.51
50-NN	WTVA	91.2	10.43	1.02
	HTVA	90.83	9.28	0.97

由表 1 可以看出，HTVA-File 方法的过滤效率要略高于 WTVA-File 方法。本文还实现基于第一维分量排序的 HTVA-File 方法(SHTVA)，表 2 给出了该方法的访问近似向量比例和第一阶段过滤效率。

表 2 SHTVA 方法性能

	访问近似向量比例(百分比)			第一阶段过滤效率(百分比)		
	2bits	4bits	6bits	2bits	4bits	6bits
10-NN	92.72	24.56	19.87	89.31	2.98	0.31
20-NN	92.83	25.69	21.32	90.02	4.06	0.42
50-NN	92.98	26.03	22.18	90.64	7.31	0.79

由表 2 可知，SHTVA 方法从第一维分量匹配的近似向量开始访问近似向量文件，进一步提高了第一阶段的过滤效率。更为重要的是，SHTVA 方法减少了近似向量的访问数量，当采用 4 位/维的平均位串长度时，只需要访问 1/4 左右的近似向量即可。这样，SHTVA 方法达到了降低 I/O 复杂度和查询时间的双重目的，能够显著提高 VA-File 方法在大型高维图像数据库中的应用能力。

结论 针对大规模高维图像数据库检索，提出一种 Hadamard 变换域的向量近似方法。与小波变换相比，Hadamard 变换的能量集中特性更好，而且运算简单。目前的变换域向量近似方法没有充分利用变换域向量的能量分布特性。笔者在能量最大的分量上建立顺序索引，通过低维过滤算法可以排除大量的不匹配近似向量，即减少了 I/O 访问时间，又提高了近邻查询效率。实验结果表明，本文提出的方法显著提高了变换域向量近似方法的性能，适合对大规模高维数据库进行近邻查询。

参 考 文 献

1 Rui Y, Huang T S, Chang S F. Image Retrieval: Current Techniques, Promising Directions, and Open Issues [J]. Journal of Visual Communication and Image Representation, 1999(10): 36～92

2 Bohm C, Berchtold S, Keim D. Searching in High-Dimensional Spaces-Index Structures for Improving the Performance of Multimedia Databases [J]. ACM Computing Surveys, 2001, 33(3): 322～373

3 Weber R, Schek H J, Blott S. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces [A]. In: Proc. 24th Int. Conf. VLDB [C], New York, IEEE, 1998. 194～205

4 Kanth K V, Agrawal D, Singh A. Dimensionality reduction for similarity searching in dynamic databases [A]. In: Proc. ACM SIGMOD Int. Conf. on Management of data [C]. Seattle, Washington, 1998. 166～176

5 Indyk P, Motwani R. Approximate nearest neighbors: Towards removing the curse of dimensionality [A]. In: Proc. 30th ACM Symposium on Theory of Computing [C], Ne York, ACM, 1998. 604～613

6 Ferhatosmanoglu H, Tuncel E, Agrawal D. Vector approximation based indexing for non-uniform high dimensional data sets. In: Proc. of the ACM Int'l Conf. onInformation and Knowledge Management(CIKM2000). New York: ACM, 2000. 202～209

7 崔江涛，孙君顶，周利华. 基于小波变换的多分辨率高维图像检索方法. 西安电子科技大学学报，2005，32(3)：370～373

8 Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data. IEEE Trans on Pattern Analysis Machine Intelligence, 1996, 18(8): 837～842