文章编号:1004 - 1478(2005)04 - 0042 - 03

基于内容的多媒体数据库索引算法研究

徐 焕[1], 林坤辉[2]

(1. 厦门大学 软件学院, 福建 厦门 361005;
2. 厦门大学 计算机科学系, 福建 厦门 361005)

摘要:分析了多媒体数据库信息检索的索引算法,并且对 k-means 聚类算法中初始聚类数目和聚类中心的设定进行了改进,设计了一种用于大容量图像数据库的索引方法. 在 1 万多幅风景图像数据库上反复进行实验,结果表明该算法能够有效地支持大容量图像数据库的基于内容的检索.

关键词:多媒体数据库;索引算法;多维索引;聚类
中图分类号:TP391 **文献标识码**:A

Research of indexing algorithm on the content-based multimedia database

XU Huan[1], LIN Kun-hui[2]

(1. *College of Sofware, Xiamen Univ., Xiamen* 361005, *China*;
2. *Dept. of Comp. Sci., Xiamen Univ., Xiamen* 361005, *China*)

Abstract: With study on the indexing algorithm of multimedia database information retrieval and the method of setting the initial number and the centers of the k-means clustering, an improved indexing method is designed for large database. Experiments on more than ten thousand scenery image database show this indexing method is effective for content-based retrieval in large image database.

Key words: multimedia database; indexing algorithm; multi-dimensional indexing; clustering

0 引言

多媒体数据库内容丰富,有文本、图像、视频等. 在基于内容特征的检索中,特征矢量高达 100 多级量,大大高于常规数据库的索引能力,因此,需要研究新的索引结构和算法,以支持快速检索.

高效的索引技术是基于内容的检索在大型数据库中发挥优势的保证. 索引技术随着数据库的发展而发展,提高索引效率有缩减特征向量的维度和改进索引算法两种方法,针对图像检索需要 3 个步骤: 1) 进行维度约减; 2) 对存在的索引方法进行评价; 3) 根据评价定制自己的索引方式. 目前多维索引技术研究较多的是聚类和神经网络. 聚类就是按照一定的要求和规律对事物进行区分和分类的过程,在图像数据库中,聚类就是在研究大量图像特征的基础上通过学习产生出类别,然后按次类别对图像进行分类. 它的优势就是可以动态地进行图像分类,而且可以有效地降低维度和查询范围,提高查询效率. 聚类的方法也有很多,最常用的是 k - means 算法,这个算法简单、有效,但要先确定类的数目,即初始类别数和初始聚类中心要预先设定,这些初始参数

收稿日期:2005 - 05 - 20

作者简介:徐焕(1975 —),女,河南省禹州市人,厦门大学硕士研究生,主要研究方向:网络多媒体、数据库应用.

将直接影响最后的聚类结果.很多情况下,这些参数都是未知的,要靠人的主观经验输入,往往与客观实际有很大误差.针对 k - means 算法的不足,笔者设计了一种算法,改进了 k - means 算法的初始模板的设定方法,使得聚类个数和聚类中心由数据分布的统计密度动态生成,无需人的参与,这样,最后的聚类结果更符合客观实际.

1　现有的聚类索引算法

在多媒体数据基于内容的索引算法中,常用聚类方法来降低维度和缩小查询范围.现有的聚类算法多是为模式识别而设计的,将目标用其特征来表示,一个目标表达为多维特征空间的一个点,在特征空间中聚类.常用的聚类算法有分割算法、层次算法、基于密度的方法、基于网格的方法和基于模型的方法[1].分割算法是将 n 个目标划分到 k 个聚类中去,k 为输入的参数.首先选择 k 个代表点,其余目标根据到各类代表点的距离划分到 k 个聚类中;然后用每个类的中心(k - means 算法)或离中心最近的点(k - medoid 算法)代表这个聚类,将目标重新分割;这一过程迭代进行,直至收敛.分割算法适用于聚类为凸形状和各类相距较远且直径相差不多的情况,否则可能产生错误的分割.层次算法将数据集分解成树状图,即循环地将数据集分裂成子集,直到每个子集只包含一个目标.树状图可采用分裂或合并的方法构建.层次算法不像分割算法那样需要聚类数这个参数,但需要定义停止条件.层次算法的难点在于最优停止条件难以确定,同时也难以处理聚类形状复杂的情况.

2　k - means 算法分析与实现原理

2.1　k - means 算法分析

k - means 算法是所有聚类算法中应用最广泛的一种分割而非分层的聚类方法.它的基本思想为:给定一个例子集合 n 和一个整数 k,k - means 算法将 n 分割为 k 个聚类并使得在每个聚类中所有值与该聚类中心的距离总和最小,其中聚类中心是指该聚类的几何平均值.该算法具有以下优点:能有效地处理大数据集;迭代速度快;算法的执行结果往往与初始模板的选定以及事例的顺序有关;算法终止于一个局部最优解.其缺点在于要设定初始模板,k 值如果不合适会导致聚类不合理.

2.2　算法的实现原理

设 n 为图像库中的图像数目.首先将图像库中的图像聚为 k 类,每个类中都有一个代表样本,即聚类中心.每类中平均有 n/k 幅图像,但是具体每个类中的图像数目与图像的颜色特征分布有关.通常是取 $\max(\sqrt{n}, d)$ 来确定 k(其中 d 为特征维数).

在进行检索时,先将示例图像与每个类的聚类中心进行相似度匹配,找到与示例图像距离最近的聚类中心.然后在该聚类中,将示例图像与每一幅图像进行匹配,根据用户所需的相似度要求,将查询结果排序并返回.在对图像库中的图像进行聚类处理后,库中所有图像根据与聚类中心距离的远近程度,形成 k 个互不相交的聚类,较为相似的图像都聚在同一个类中.因此,示例图像只需与各聚类中心相比较,再在最相近的类中进行匹配,即可得到较好的查询结果.该算法的平均匹配次数为 $p = k + n/k$.由于 $k < n$,因而相对于顺序查找的匹配次数,算法的匹配次数显然是大大减少了.相应地,查询时间也会因此而减少,查询效率则大为提高.

算法实现的具体步骤如下:

1)给定分类个数 k 以及初始 k 个聚类中心 Z_1, Z_2, Z_3, …, Z_k;

2)对于图像库中每幅图像,计算其与各聚类中心的距离,将之归入距离值最小的类中;

$X \in C_j(k)$ if $\| X - C_j(k) \| < \| X - C_i(k) \|$ for all $i = 1,2,\cdots, k; i \neq j$; where $C_j(k)$ denotes the set of samples whose cluster centre is $z_j(k)$.

3)计算新的聚类中心 $Z_j(k+1)$, $j = 1,2,3,\cdots, k$.

对于每个聚类,类中所有图像与新的聚类中心的距离和最小.

$$Z_j(k+1) = \frac{1}{N_j} \sum_{x \in C_j(k)} x, \quad j = 1,2,\cdots,k$$

其中 N_j 是每个类中图像的数目.

4)如果 $Z_j(k+1) = Z_j(k)$, $j = 1,2,3,\cdots, k$,那么算法是收敛的,程序结束;否则,返回步骤 2).

5)对每个聚类中心以及类中的图像排序,建立线形链表,程序结束.

3　k - means 算法的改进与验证

3.1　k - means 算法的改进[4,5]

对 k - means 算法的改进,主要是改进了初始模板的选定方法.以每个图像为圆心,以数据库中所有图像之间距离的平均值为半径作圆,然后根据每个圆内的数据点的密度来排序确定初始聚类中心和初始聚类数.这样,k - means 聚类算法需要的初始模板就由以上算法动态生成,而无需用户进行事先指定.整个过程包括以下几个基本步骤:

1)将数据库中的每幅图像都看成一个类,计算所有图像之间的距离,生成距离矩阵;

2)选取 2 个正数,一般 $R_2 = 2R_1$,其中 R_1 为数据库中所有图像之间距离的平均值;

3)以每幅图像为圆心,以 R_1 为半径作圆,计算落在每个圆内的图像数目,即样本密度;

4)将样本密度按从大到小的顺序排列,取密度最大者作为第一个凝聚点 Z_1,在密度次大的单元中任选一点 k,若与第一凝聚点之间距离大于 R_2,即 $|Z_1 - k| \gg R_2$,则把 k 作为第二个凝聚点 Z_2,否则继续判定下一密度最大者,若下一密度最大者中的任一点与前面若干个凝聚点之间距离均大于 R_2,则将之作为又一新的凝聚点,如此反复迭代直到没有新的凝聚点生成;

5)这些凝聚点作为聚类模板的初值即分类个数 k 以及初始 k 个聚类中心 $Z_1, Z_2, Z_3, \cdots, Z_k$;

6)把得到的 k 和 k 个聚类中心 $Z_1, Z_2, Z_3, \cdots, Z_k$ 作为 k - means 算法的初始模板,继续用 k - means 算法迭代,最后得到 k 个聚类.

经过初始分类,可以得到整个数据库的分类个数 k 以及模板初始聚类中心 $Z = \{Z_1, Z_2, Z_3, \cdots, Z_R\}$,然后,进行 k - means 迭代. 其基本原理是根据图像数据库中所有图像与聚类中心距离的远近程度,形成 k 个互不相交的聚类,较为相似的图像都聚在同一类中.

在整个索引算法中,聚类的过程最为关键,聚类的效果将直接影响查询结果的优劣. 改进后的聚类算法是动态的,初始模板的设定不再依赖人的经验参数,而是从整个库的统计特性中获取必要的参数信息. 因而最后的聚类结果更加客观合理.

3.2 对改进的 k - means 算法的验证

为了检验算法的有效性,建立了测试模型,测试在 1 万多张花卉和山水风景图像数据库上进行. 抽取图像的颜色特征,以相同的测试条件(即相同的示例图像和相同的相似度),分别用顺序查找、k - means 算法和改进的 k - means 算法测试,检查返回结果的时间和返回的准确性,部分测试数据显示如表 1.

表 1 1 万张风景图像上的测试结果

算法	查到的数	查找时间/ ms	查到率
顺序查找	368	7.60	100 %
k - means 算法	300	4	81.5 %
改进的 k - means 算法	180	1.03	48.9 %

注:测试条件:库中图像数目 = 10060 幅,查找示例:山水风景图像,相似度 = 50 %.

对于大容量图像库而言,很难找到与示例相关的所有图像数目,因为"相关"的概念是模糊的,人们由肉眼感受到的相关机器却很难识别. 不妨认为顺序查找的查到率是 100 %,其他算法的查到率定义为:查到率 = 查到的数/顺序查找查到的数. 以上测试表明,改进后的 k - means 算法采用的是动态生成的初始参数,大大减少了 k - means 算法的迭代次数,而且在聚类基础上建立的索引结构,检索时访问的不相关节点数也大大减少,节约了不少访问这些节点的开销,因而极大地缩短了查询时间,几乎是顺序查找的 1/7,而且查到的不相关图像大大减少,这个结果基本能够满足用户对大容量数据库检索的需求.

4 结论

通过对各种聚类算法进行分析、比较和实验,提出了一种改进的 k - means 算法. 该算法可满足大容量数据库检索的需要,但也存在一些有待改进的问题:其一,随着相似度的降低,索引效果更为明显,与此同时,查到率也随之降低. 可见,图像的查到率与查找时间是互为矛盾的. 要想减少查找时间,需要以降低查到率作为代价. 因此,应该设法寻找二者的性能最优点,使用户的需求尽可能地得到满足. 其二,文中所提出的索引算法主要是针对图像数据的低层物理特征,用的是颜色特征来进行测试的. 笔者对于形状特征也进行了实验,但效果不及颜色特征好. 因此,需要进一步研究针对多种特征的索引建立方法,以便更好地满足大容量多媒体信息的快速查询. 其三,用户在进行检索时,有可能要求同时进行关键字检索和基于内容的检索. 传统数据库采用的是基于关键字的索引方法,而多媒体数据库采用的是高维特征的索引方法,二者之间存在很大的差别,基于内容的检索系统并不排斥那些常规的检索途径,相反,要充分利用现有的文本检索功能并集成到基于内容的检索系统中,那么如何将各种索引结构结合起来,以适应不同用户的查询需求,将是未来的一个研究方向.

参考文献:

[1] 庄越挺,潘云鹤,吴飞,等. 网上多媒体信息组织与检索[M]. 北京:清华大学出版社,2003.

[2] Faloutsos C, Lin King-Ip (David). FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets[C]. San Jose, CA: ACM SIGMOD Intl. Conf. Management of Data, 1995. 163—179.

[3] Chandrasekaran S, Manjunath B S, Wang Y F, et al. An eigenspace update algorithm for image analysis[J]. CVGIP: Graphic Models and Image Processing, 1997, 59(5): 321—332.

[4] Lozano J A, Pena J M, Larranaga P. An empirical comparison of four initialization methods for the k-means algorithm[J]. Pattern Recognition Letters, 1999, 20(10): 1027—1040.

[5] A1-Daoud M B, Roberts S A. New methods for the initialization of clusters[J]. Pattern Recognition Letters, 1996, 17(5): 451—455.