基于内容的多媒体数据库管理系统的研究

李松涛，钟建宁
（广东医学院 数学与计算机科学教研室，广东 湛江 524023）

摘 要：通过分析数据库的数据模型，研究基于内容的多媒体数据库管理系统的构建方法及其功能框架，采用PL/SQL方式访问Oracle 8i数据库。为了提高管理系统的图像检索速度，提出了一种基于内容的图像检索算法，从聚类中心初值选取和分类中心的更新方面改进C-均值聚类算法，较好地解决了图像的分类问题。实验表明：使用该聚类检索算法，能对分类中心进行快速更新，有效地对图像进行聚类以及缩短检索时间，检索性能优于现有的C-均值聚类算法。

关键词：多媒体数据库；数据模型；检索；C-均值聚类；算法
中图分类号：TP315 **文献标识码**：A **文章编号**：1673-629X(2008)12-0214-03

Research of Multimedia Database Management System Based on Content Retrieval

LI Song-tao, ZHONG Jian-ning
(Dean of Computer Science and Mathematics, Guangdong Medical College, Zhanjiang 524023, China)

Abstract: Through analyzing the data model of database, studies the construction method and its functional framework of the content-based multimedia database management system by means of accessing the Oracle 8i database through PL/SQL. To increase the speed of image retrieval of the management system, proposes a content-based image retrieval algorithm which aims at improving the C-means clustering algorithm by selecting initial value of the cluster center and updating the classified center. Results of the experiment show that the application of this clustering retrieval algorithm can update the classified center quickly, cluster the relevant images effectively, shorten the retrieval time, and its retrieval performance is superior to the existing C-means clustering algorithm.

Key words: multimedia database; data model; retrieval; C-means cluster; algorithm

0 引 言

随着信息的数字化和多媒体技术的迅速发展，许多信息和数据都以多媒体数字数据的形式表现和存储。多媒体数据主要包含数字、字符、文本、声音、图形、图像和视频等形式的数据。数字和字符等有格式数据利用数据库技术进行存取管理，文本、声音、图像等无格式数据在计算机中大多以文件形式存放，由操作系统进行管理。无格式数据的特征难以用符号进行充分表达，如音频中的音调、图形的轮廓、图像中的纹理等，对这些数据进行存储、处理和查询相对就困难许多，这从而促进了多媒体数据库技术的发展。

1 数据模型

数据模型是数据库的描述机制，从不同的角度和级别描述数据库结构和信息组织方式，可见，建立数据模型是实现多媒体数据库的关键。目前实现多媒体数据管理的数据模型大致可分为三类：关系数据模型、面向对象数据模型以及超文本数据模型。

1.1 关系数据模型

关系数据模型基于关系运算理论和关系模式设计理论，由关系数据结构、关系数据操作和关系完整性约束三部分组成。基于关系数据模型的数据库系统有效解决了有格式数据的许多管理问题，但对于复杂的多媒体信息处理则显得困难。因此，对原有的关系数据库加以扩充，使之能支持多媒体信息的处理，如从Oracle 8i开始，增加了LOB型字段用于多媒体数据等大型对象的存取[1]。虽然关系数据库系统只能处理有格式数据，语义表示能力差，但有良好的兼容性和广泛的应用基础，通过生产商对其功能的增加和改进，能够存储和管理多媒体数据并实现对数据的查询和检索。

1.2 面向对象数据模型

面向对象的方法通过对象结构、对象表示、封装、

收稿日期：2008-03-08
基金项目：2006年度湛江市科技攻关项目(2006C09018)
作者简介：李松涛(1975-)，男，广东湛江人，讲师，硕士，研究方向为数据库、信息系统。

继承等方式能够有效地描述各种对象及其内部结构和联系,适用于描述复杂对象。面向对象数据模型是由类构成的层次结构,类是对象的抽象,类与类之间的继承关系构成类的层次结构能够完整描述现实中的数据结构,语义表达能力丰富,能够充分反映和管理多媒体数据的特征以及各种媒体数据之间的时间和空间的关联,但模型比较复杂。

1.3 超文本数据模型

超文本是由节点和表达节点之间的链组成的网,以非线性方式组织内容。节点描述超文本数据对象的内容,链定义超文本的结构。超文本提供了沿链访问数据的方法,符合人类联想式的思维习惯,是一种高级的数据库技术。超媒体是超文本的结点与链推广到多媒体的形式,用于表示、组织、存储、访问多媒体文档,是目前能够较好描述多媒体数据的模型。

2 基于内容的多媒体数据库管理系统的关键技术

基于内容的多媒体数据库管理系统除了解决媒体信息的存储,还要求能够从媒体数据中分析、抽取可供检索的内容特征,用于信息的检索[2]。

2.1 特征抽取

特征抽取为基于内容的检索提供基础支持,可以人工操作完成,也可以通过对应的媒体处理程序完成。不同的媒体信息的描述需要抽取不同的特征,如图像有颜色、形状、纹理等特征;音频有音调、亮度、响度、带宽等特征;视频的主要特征则是代表帧。抽取的特征可以是全局性的,也可以针对某个内部的对象。

2.2 颜色检索

颜色是图像的一个重要特征,具有一定的稳定性,可用 RGB 三色表示法、HSI 表示法、CIE 的色度和亮度表示法等进行描述。在研究初期颜色通常作为图像特征应用于图像检索,其步骤是先选择合适的颜色空间来描述颜色特征,然后通过量化将颜色特征转化为向量的形式,最后定义相似度标准衡量图像之间在颜色上的相似性[3,4]。

2.3 纹理检索

纹理特征是图像局部性质(灰度分布函数)的统计。纹理特征包含了物体表面结构组织排列的重要信息以及它们与周围环境的联系,适合于描述山脉、云彩、树木、纤维等图像。通过结构分析方法和统计分析方法可以测量纹理特征[5]。

2.4 视频检索

通过视频分割可将视频分成检索单元,其中包括场景分割和镜头分割。将视频中某个有意义的情节独立为一个检索单位称为场景分割;镜头是视频检索的基本结构单元,将其独立出来的过程称为镜头分割。实际操作中计算镜头往往通过在场景变换中提取关键帧来完成。关键帧是镜头的一种简单有效的表达,是镜头中最有代表性的一幅或多幅图像。在基于内容的视频检索中,关键帧既可作为静态图像处理,也可用于视频浏览。

2.5 音频检索

音频数据有着物理和感知两类特征。物理特征来源于音频信号本身,包括短时能量、频率等;感知特征依赖于人的听觉模型,包括音调、音高等,根据这些特征进行分类能够检索出所需的音频。为方便视频检索的进行,可根据音频特征对音频数据进行分段。

3 基于内容的多媒体数据库管理系统功能框架的设计

基于内容多媒体数据库管理系统的的功能框架如图 1 所示。

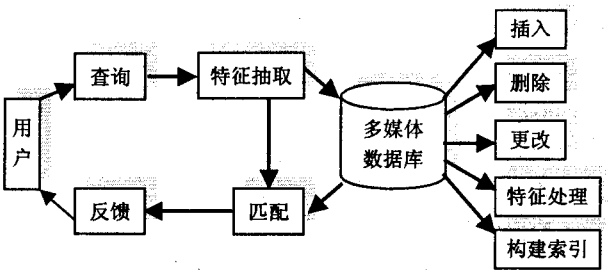

图 1　系统功能

(1)查询模块是为用户提供符合需求的多媒体信息。

(2)特征抽取模块是分析用户给定的样式信息,对信息的特征进行提取和描述。

(3)匹配模块是将提取后描述的特征与多媒体数据库中各样本的特征进行比较,为用户选取最相近的多媒体信息。

(4)用户反馈模块是让用户根据自己的要求,对搜索得到的多媒体数据进行判断,然后确定是否需要修改先前提交的查询。

对数据库的操作主要是日常维护工作(如插入、删除、更改等)和多媒体数据的特征处理。多媒体数据经过分析处理后进行特征抽取,以特征向量的形式存储,由于特征向量的维度较高,需要用多维索引算法为特征向量构建索引。

结合成本和技术成熟程度等因素,基于内容的多媒体数据库管理系统以关系模型为基础,添加面向对象层,结合为对象—关系数据库管理系统,支持多媒体信息。Oracle 8i 通过 LOB 型字段支持多媒体信息,可

细分为二进制大数据对象(BLOB)、字符大数据对象(CLOB)、民族语言化字符大数据对象(NCLOB)、文件大数据对象(BFILE)。由于大数据对象的实际大小很难固定且有一定的特殊性,因此需要通过过程模块的调用进行存储。在 Oracle 中支持 LOB 操作的工具有 OCI 和 PL/SQL,由于 OCI 的编程接口比较复杂,而 PL/SOL 是 Oracle 对关系数据库语言 SQL 的过程化语言扩充,所以采用 PL/SOL 来处理多媒体数据。

4 聚类算法

聚类是一个将数据集合划分为若干组或类的过程,相似的对象聚集在同一类中,所以同组内的数据具有较高的相似性,而不同组的数据是不相似的。在多媒体数据库中基于内容的检索索引可以通过聚类算法实现。

C-均值聚类算法是一种广泛应用的聚类算法,使用时要预定类的数量,聚类效果易受初始聚类等因素的影响[6]。而模糊 C-均值算法利用伪随机数产生初始的类中心,聚类效果因随机数的选取变得不稳定。本系统对 C-均值聚类算法进行改进,解决特征向量与质心距离相等时的样本分配问题,加快聚类速度,对图像库的聚类处理有较好效果。

改进的 C-均值聚类算法步骤如下:

1)确定初始类数目和类中心。

选取比较合理的初始类数量和类中心是聚类算法中关键的第一步。设定数据库中的图像对象数目为 T,图像的特征向量维数为 N,则数据库中的所有图像可表示为 T 个 N 维特征向量。初始类的数量 K 根据式(1)得到[7],因为由随机函数的分布可知聚类的数据主要是分布在全体数据的均值周围,而在数据处理过程中应用几何平均数可表示集中趋势。

$$K = \max(\sqrt{T}, N) \tag{1}$$

从特征库中选取距离最大的两个特征向量 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 分别作为第一、第二个初始类中心,根据式(2)计算其它 $K-2$ 个初始类中心。如果已确定 K 个类,则第 $K+1$ 个类是与前 K 个类最不相同的,也就是第 $K+1$ 个类的中心是与前 K 个类的中心距离累加和最大的特征向量[8]。

$$\sum_{i=1}^{K} D(\boldsymbol{X}_i, \boldsymbol{X}_{K+1}) = \mathrm{MAX}\{\sum_{i=1}^{K} D(\boldsymbol{X}_i, \boldsymbol{X}_j), j = 1,2,\cdots, T-K\} \tag{2}$$

$D(\boldsymbol{X}_i, \boldsymbol{X}_j)$ 为向量 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 之间的欧氏距离

$$C_i = \boldsymbol{X}_i \quad C_i \text{ 为第 } i \text{ 类的质心} \tag{3}$$

2)样本分配。

根据式(4)计算每一个对象样本与类中心的欧氏距离,将其欧氏距离最小的样本分配到对应中心的类。

$$D(\boldsymbol{X}_m, C_k) = \mathrm{MIN}(D(\boldsymbol{X}_t, C_k), t = 1,2,3,\cdots, T) \tag{4}$$

$\boldsymbol{X}_m \in \mathrm{genus}[k]$ $\mathrm{genus}[k]$ 代表第 k 类

3)更新聚类中心。

根据式(5)计算各类中成员的质心,调整聚类中心,重新进行对象样本的分配,重复进行直到各个聚类的中心不再变化。

$$C_k = \frac{1}{\mathrm{num}[k]} \sum_{\mathrm{genus}[t] \in k} \boldsymbol{X}_t \quad t = 1,2,3,\cdots, T \quad k = 1,2,3,\cdots, K \tag{5}$$

$\mathrm{num}[k]$ 为第 k 类中对象样本的个数。

C-均值算法根据欧氏距离分配样本,模糊 C-均值算法根据特征向量之间的距离计算模糊权,如果某一向量与两个类的质心距离相等,样本不能得到明确的分类。用 gaussian 模糊权能够解决这个问题,但其模糊权的计算繁杂耗时,此处提出质心向量差方法,使这个问题得到较好解决。

设定 $C_k = [c_1, c_2, \cdots, c_n]$,当不能确定某个样本归于第 A 类还是第 B 类时,根据式(6)将两类的质心相减,当 $a_i - b_i$ 的差不等于零,可以作出判断,如果差大于零该样本归于第 A 类,若小于零则归于第 B 类。因为两类的质心是不同的,肯定会出现 $a_i - b_i$ 的差不等于零,这样可将样本确定在某一个类中。在抽取样本特征向量时要确定特征分量对样本的影响权重,根据权重大排列在前面的原则确定特征分量的排列次序,这样能有效地快速解决样本的分类问题。通过实验,改进的 C-均值算法速度较快。

$$A - B = [a_1 - b_1, a_2 - b_2, \cdots, a_n - b_n] \tag{6}$$

5 结束语

基于内容的多媒体数据库技术是目前数据库研究的热点。文中从数据模型、功能框架和检索技术中的聚类算法等方面探讨了如何构建多媒体数据库管理系统及其存储检索问题。对 C-均值聚类算法中初始类数量的确定和样本的分配作了改进,有效地解决了图像检索中分类的速度问题。

参考文献:

[1] 黄金敢,潘 敏.基于文件名的多媒体数据库管理系统开发[J].计算机工程与设计,2005,26(4):1048-1050.

[2] 邱建雄,赵跃龙,李国辉.基于聚类的图像视觉内容检索和索引[J].小型微型计算机系统,2005,26(3):492-495.

[3] 陈峰莲,阎保平,黎建辉,等.科学数据库基于内容的多媒体检索系统[J].计算机科学,2005,32(2):97-99.

(下转第 223 页)

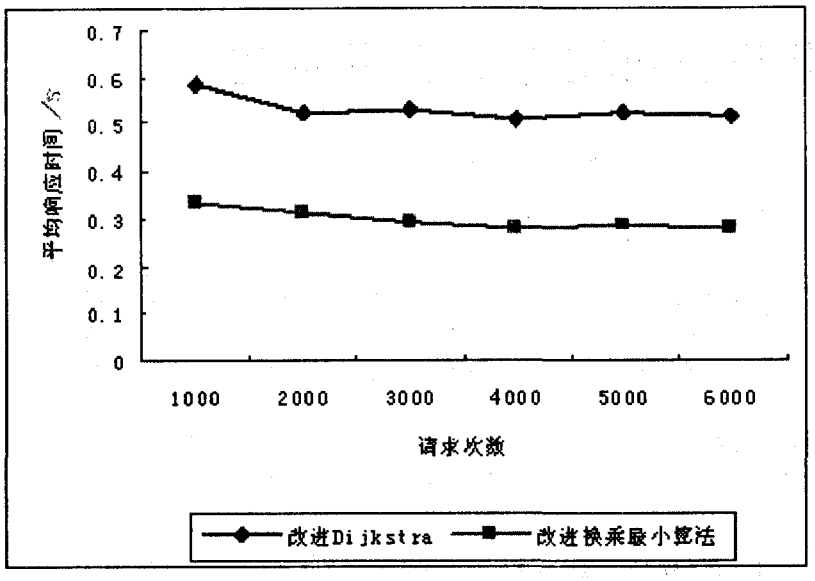

图 4　采用改进 Dijkstra 和改进的换乘最小算法平均响应时间图

算法对本模型进行验证,并成功运用在智能交通信息平台测试版中,取得了较好的效果,文中通过公交乘车的宏观因素来分析问题,对以后相应的公交乘车系统的开发具有一定的实用意义。但该模型同样也存在需要改进的地方,如对实时公交路况等的支持上,这也将成为今后工作的重点。

参考文献:

[1] 张林峰,范炳全.公交网络换乘矩阵的分析与算法[J].系统工程,2003,21(6):92-96.

[2] 苏　啸,曾子维.基于关联的城市公交换乘查询算法[J].计算机工程与设计,2006,27(3):519-521.

[3] 傅冬绵.交通系统中最少换乘算法及其实现[J].华侨大学学报:自然科学版,2002,22(4):348-350.

[4] 王　莉,李文权.公共交通系统最佳路径算法[J].东南大学学报:自然科学版,2004,34(2):264-267.

[5] 杨新苗,王　炜,马文腾.基于 GIS 的公交乘客出行路径选择模型[J].东南大学学报:自然科学版,2000,30(6):87-91.

[6] 徐萃薇,孙绳武.计算方法引论[M].北京:高等教育出版社,2002:67-68.

[7] 张维明,邓　苏.信息系统建模技术与应用[M].北京:电子工业出版社,1997:234-235.

[8] Nievergelt J, Hinterberger H, Sevcik K C. The Grid File: An Adaptable, Symmetric Multikey File Structure [J]. Acm Transactions On Database Systems, 1984, 9(1): 38-71.

[9] Guttman A. R-trees: a dynamic index structure for spatial searching[C]//proceedings of the 1984 ACM SIGMOD international conference on Management of data. New York: ACM Press, 1984: 47-57.

[10] 叶　青,陈闵中.基于预处理剪枝的最短路径算法[J].计算机工程,2007,43(9):205-207.

(上接第 216 页)

[4] 郭　健,陈孝威.基于颜色分布的图像检索技术[J].计算机工程与应用,2006(6):44-47.

[5] 王永良,陈新度,刘　强,等.基于内容的墙纸 Web 检索系统的研究与实现[J].计算机应用研究,2006,23(6):167-169.

[6] 刘　笛,朱学峰,苏彩红.一种新型的模糊 C 均值聚类初始化方法[J].计算机仿真,2004,21(11):148-151.

[7] Kim Tae·Wan, Li Ki-Joune. A distance based packing method for high dimensional data[C]//Proceedings of the Fourteenth Australasian database conference on Database technologies. Adelaide, Australia: [s. n.], 2003: 135-144.

[8] 张培珍,付　平,肖　军,等.基于快速聚类索引的图像检索系统[J].吉林大学学报:信息科学版,2004,22(6):638-642.

(上接第 219 页)

schemes for IP traceback[C]//In: Proceedings of the 2001 IEEE INFOCOM Conference. Anchorage, Alaska: [s. n.], 2001.

[2] Carter R, Crovella M. Dynamic Server Selection Using Dynamic Path Characterization in Wide-Area Networks[C]//In: Proceedings of the 1997 IEEE INFOCOM Conference. Kobe, Japan: [s. n.], 1997.

[3] Cooperative Association for Internet Data Analysis, Skitter analysis[EB/OL]. 2000. http://www.caida.org/tools/measurement/skitter/.

[4] Savage S, Wetherall D. Karlin A, et al. Practical Network Support for IP Traceback[C]//Proceedings of the 2000 ACM SIGCOMM Conference. Stockholm, Sweden: [s. n.], 2000: 295-306.

[5] Savage S, Wetherall D, Karlin A, et al. Network support for IP traceback [J]. IEEE/ACM Transactions on Networking, 2001, 20(2): 226-237.

[6] Deering S, Hinden R. Internet Protocol Version 6 (IPv6) specification[S]. RFC 2460. 1998.