

分布式远程备份空间模型研究^{*}

魏青松¹ 陈波² 卢显良² 周旭² 吴劲²

(同济大学软件学院 上海200331)¹ (电子科技大学计算机学院 成都610054)²

摘要 本文提出一种有效的寻找条块化存储最优化策略的方法,根据分布式备份的基本需求,定义了描述条块化策略好坏的性能参数,定量给出了各种空间存储模型下的性能计算方法,从空间几何的角度用多面体形象地描述各种存储模型,并研究了四面体和六面体存储模型下各种存储策略的性能和计算方法。

关键词 分布式,备份,条块化,空间存储模型

The Research of the Distributed Distant Backup Spatial Model

WEI Qing-Song CHEN Bo LU Xian-Liang ZHOU Xu WU Jin

(School of Software Engineering, Tongji University, Shanghai 200031)¹

(Department of Computer Science of UEST of China, Chengdu 610054)²

Abstract This paper presents an efficient method for finding optimized striped storage policy, abstracts and defines performance parameters to describe striped policy according to basic requirement of distributed backup, and deduces performance evaluation formula upon several spatial storage model, describes the storage model with polyhedron, and researches the performance of various storage policy upon tetrahedron and hexahedron storage model and finds the optimized one.

Keywords Distributed, Backup, Striping, Spatial storage model

1 引言

数据的高可用性要求存储的高可靠性。分布式远程备份是保证数据可靠性的关键技术。分布式远程备份技术充分利用网络技术将数据备份到远程各节点,实现了分布式数据容错和远程灾难恢复,克服了本地备份(如 RAID 磁盘阵列、镜像盘)只将数据备份到本地的局限性^[1,2]。

分布式远程备份一般采用分布式条块化存储,图1所示,首先将数据分割成几块,将各数据块分布到各远程节点。实际部署中,数据如何条块化、节点如何选择将直接影响数据存储的整体性能。选择好的条块化存储策略,在容错性、文件访问速度、资源开销等方面获得高的综合性能是分布式远程备份的关键^[3]。

本文提出一种有效的寻找最优化存储策略的方法,根据分布式备份的基本需求,定义了描述条块化策略好坏的性能参数,定量给出了各种存储模型下的性能计算方法,从空间几何的角度用多面体形象地描述各种存储模型,并研究了四面体存储模型的各种性能参数和计算方法。

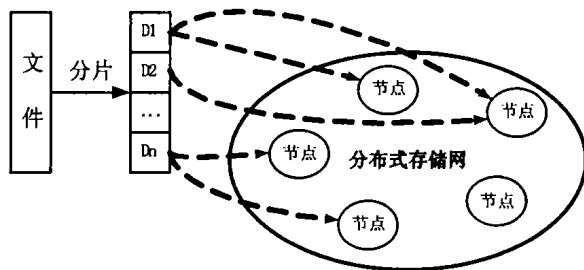


图1 分布式条块化存储体系结构

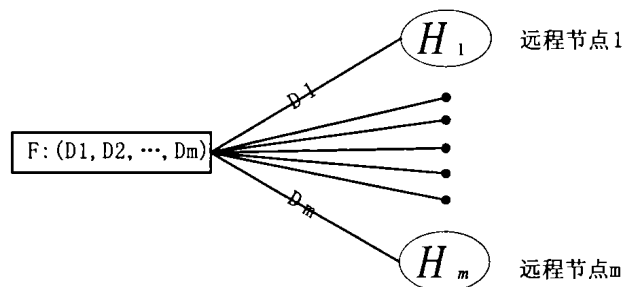


图2 一维存储模型

2 分布式条块化存储模型分析

依据分布式存储的观点^[4,5],可以将一个文件分割为 m 部分,记为 $\{D_1, D_2, \dots, D_m\}$,由 n 台主机来存储这些数据块,其中任意一台主机 $H_i (i \leq n)$ 备份 m_i 块数据,记为 $\{D_{i,1}, D_{i,2}, \dots, D_{i,m_i}\}$,它们都是 $\{D_1, D_2, \dots, D_m\}$ 的子集。本文提出以下性能指标,作为衡量条块化策略好坏的标准。

备份可靠性: 某些存储节点出现故障时,仍可通过剩余的节点恢复文件的概率,简记为 R (Reliability)。

节点依赖性: 可以恢复文件所需的最少主机的平均数/主机总数,简记为 D (Dependency)。

文件分解度: 文件被分割为多少个数据片,简记为 P (Parts)。

空间冗余度: 每个数据块所拥有备份的平均数,简记为 S (Space)。

备份可靠性是基本性能参数,其值 (R) 越高,说明文件越容易被恢复;节点依赖性描述文件恢复的困难程度,在什么样

^{*} 本文受国家95重点攻关项目支持。魏青松 博士研究生,主要研究方向:计算机网络、网络存储、分布式操作系统等。陈波 副教授,卢显良 教授,博士生导师,主要研究方向:计算机网络、操作系统。

的边缘条件下,文件仍然可恢复,其值(D)越小,说明相对于整个节点总数来说,只需要很少的远程节点就可以完成文件的恢复;文件分解度(P)越大,说明文件分得越细,获取整个文件的代价越高,同时也可获得较高的并行传输带宽;空间冗余度(S)越大,说明数据块在远程主机有足够的拷贝,因而文件越有可能被恢复,同时也占用更多的磁盘空间。

2.1 一维存储模型

先讨论简单情况,将文件分为 m 个数据块,记为 $\{D_1, D_2, \dots, D_m\}$, 每块只备份在一台远程备份主机上,而且每台主机只存储一块数据,那么需要 m 台节点主机,很显然这是一种一一映射,如图2所示。

如果用图的概念来解释,那么就是 m 个孤立的节点 $\{H_1, H_2, \dots, H_m\}$, 每个节点 H_i 保存数据块 D_i , 可以称这种情况为一维存储模型,相应地,一维存储模型的性能计算方法称为一维算法。

假设每台主机出现故障的概率为 P_i , 那么要恢复文件 F 必须在所有的节点都可用的情况下才能完成。所以备份可靠性 R 为:

$$R = P(H_1 * H_2 * \dots * H_m)$$

由于这 m 台主机之间相互独立,因此有:

$$R = P(H_1) * P(H_2) * \dots * P(H_m)$$

因此,代入 P_i 可以得到备份可靠性 R :

$$R = \prod_{i=1}^m (1 - p_i)$$

文件被分解为 m 份,所以文件分解度 P 为: $P = m$

要恢复数据,必须使 m 台主机都可以用,所以节点依赖性 D 为: $D = m/m = 1$

每个数据块只有一个备份,所以空间冗余度 S 为: $S = 1$

2.2 二维存储模型

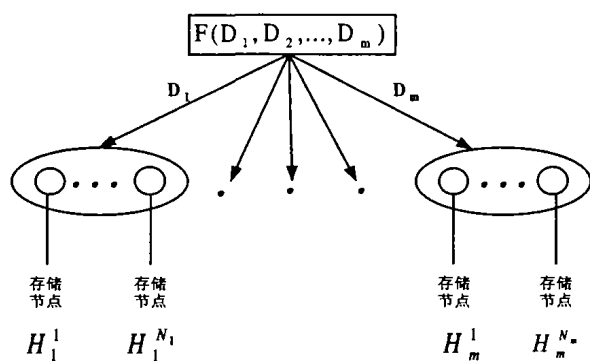


图3 二维存储模型

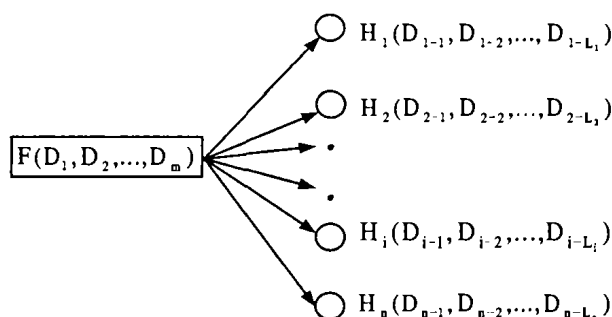


图4 多维存储模型

将文件分为 m 块,记为 $\{D_1, D_2, \dots, D_m\}$, 每块数据 D_i 不是只备份到一台远程备份主机上,而是备份到 N_i 台主机,但是每台主机只备份一块数据,这是一对多的映射关系,如图3

所示。

图3中备份 D_i 数据的第 j 台主机记为 H_i^j 。同样用图的概念来解释,可以把备份每个数据块的主机连结成一个闭合的面。如果该数据块只有一台备份,那就是一个点($\cdot$),如果是两台那么就可以连接成 $(A \text{ --- } B)$ 这样的面,如果是两台以上(例如 N_i 台, $N_i \geq 3$)就可以连接为 N_i 边形。因为出现拉面,可以称这种情况为二维存储模型,相应地,一维存储模型的性能计算方法称为一维算法。

假设 H_i^j 主机出现故障的概率为 P_i^j ; 要恢复数据块 D_i , 那么在备份数据块 D_i 的主机中至少有一台是可用的。记数据块 D_i 可用的概率为 P_i , 由于主机之间相互独立,则有:

$$P_i = P(1 - P_i^1 * P_i^2 * \dots * P_i^{N_i})$$

代入概率值可得:

$$P_i = P(1 - \prod_{j=1}^{N_i} P_i^j)$$

要恢复文件 F , 必须获得全部 m 个数据块,因此备份可靠性 R 为:

$$R = \prod_{i=1}^m P_i$$

文件被分解为 m 份,所以备份分解度 P 为: $P = m$

要恢复数据,需要 m 台主机,所以备份依赖性 D 为:

$$D = m / \sum_{i=1}^m N_i$$

由于数据块 D_i 拥有 N_i 个备份,因此空间冗余度 S 为:

$$S = \sum_{i=1}^m N_i / m$$

可见, D 与 S 呈倒数关系,和一维算法相比,在同样的分解度下,依赖性减小了,而冗余度增大了,但是需要的主机过多,原因就是每台主机只备份了一块数据,没有得到充分的利用。

2.3 多维存储模型

将文件分为 m 块,记为 $\{D_1, D_2, \dots, D_m\}$, 任一数据块 D_i 都被备份到 N_i 台主机上,但是每台主机不只备份一块数据,而是多块数据。假设主机 H_i 备份 L_i 块数据,记为 $\{D_{i-1}, D_{i-2}, \dots, D_{i-L_i}\}$, 它们都是 $\{D_1, D_2, \dots, D_m\}$ 的子集,假设共有 n 台主机 $\{H_1, H_2, \dots, H_n\}$ 来备份这些数据,这是一种多对多的映射关系,如图4所示。

结合二维算法中图的解释方法,这里两个面出现了相交的边,这种相交情况可能仍可以用平面来表现,但是很多情况下平面是不能够表现出来的,必须用三维或多维空间才能够表示。为此,称这种情况下的模型叫做多维存储模型,相应地,一维存储模型的性能计算方法称为一维算法。

为了方便计算多维情况的备份性能,提出如下定义:

恢复包: 可以恢复文件的远程备份主机集合。就是说,当这些主机可用时,文件就一定可以恢复(显然,全部主机的集合就是一个恢复包)。

恢复闭包操作: 如果在恢复包中删除某台主机,剩下的主机集合仍然是恢复包,那么,就称这一删除操作为恢复闭包操作。

恢复闭包: 不能够再进行闭包操作的恢复包,称之为恢复闭包。

不可恢复包: 不可恢复文件的远程主机集合。就是说,当这些主机出现故障,而其它的主机都可以用的情况下,文件一定不可能恢复。

不可恢复闭包操作: 如果在不可恢复包中删除某台主机,剩下的主机集合仍然是不可恢复包,那么,就称这一删除操作

为不可恢复闭包操作。

不可恢复闭包:不能够再进行不可恢复闭包操作的不可恢复包,称之为不可恢复闭包。

假如恢复闭包集合为 $C(\text{close})$, 集合 C 中的元素个数为 k , 记为 $\{C_1, C_2, \dots, C_k\}$; 而不可恢复闭包集合为 $U(\text{unclose})$, 集合 U 中的元素个数为 z , 记为 $\{U_1, U_2, \dots, U_z\}$ 。集合 C 和集合 U 都是主机集合 $\{H_1, H_2, \dots, H_n\}$ 的子集。

假设不可恢复闭包 U 中的元素 U_i 包含的主机数为 Y_i , 而且将这些主机记为 $\{H_{i-1}, H_{i-2}, \dots, H_{i-Y_i}\}$; 假设可恢复闭包 C 中的元素 C_i 包含的主机数为 X_i , 那么可以将这些主机记为 $\{H_{j-1}, H_{j-2}, \dots, H_{j-X_i}\}$; 它们都是主机集合 $\{H_1, H_2, \dots, H_n\}$ 的子集。

由不可恢复闭包定义可知, 所有不可恢复情况即是不可恢复闭包 U 中元素的并集, 因此不可恢复的概率为:

$\bar{R} = P(U_1 \cup U_2 \cup \dots \cup U_z)$ z 为不可恢复闭包 U 的元素个数

所以备份可靠性 R 为: $R = 1 - \bar{R}$

文件被分解为 m 份, 所以备份分解度 P 为: $P = m$

要恢复数据, 必须至少包含恢复闭包集合 C 的某一元素 C_i 中的主机, 而 C_i 的主机数为 X_i ; 所以, 可以恢复文件所需的最少主机的平均数为:

$(\sum_{j=1}^k X_j) / k$ k 为恢复闭包集合 C 中的元素个数

而主机总数为 n , 所以备份依赖性 D 为:

$D = ((\sum_{j=1}^k X_j) / k) / n$

由于数据块 D_i 拥有 N_i 个备份, 因此空间冗余度 S 为:

$S = (\sum_{i=1}^m N_i) / m$

3 实际分析

备份的策略相当复杂和繁多, 如何选择一个比较好的方案是必须要解决的问题。

现在来看这样一个多维情况的例子, 如图5。文件被分割为3块 (D_1, D_2, D_3), D_1 由主机 (A, B, B_1, A_1) 备份, D_2 由主机 (B_1, B_2, A_2, A_1) 备份, D_3 由 (B_2, B, A, A_2) 备份, 将其用几何图形形象化得到图5, 图5A 为其二维平面视图, 图5B 为其三维立体视图。图5B 中, 点对应的是远程备份主机, 面对应的是数据块。下面讨论四面体、六面体存储模型下的性能, 其它同理可推。

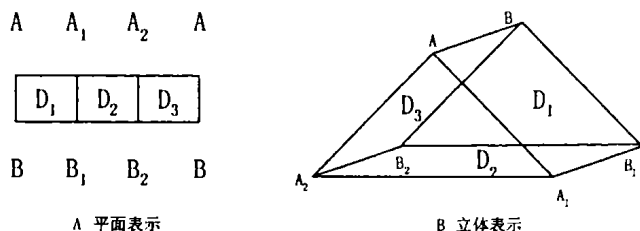


图5 多维存储的几何表示

3.1 四面体

如图6所示的四面体 $ABCD$ 。将文件分解为4份, 设 D_1 备份在主机 (A, B, C) 中, D_2 备份在主机 (A, C, D) 中, D_3 备份在主机 (A, B, D) 中, D_4 备份在主机 (B, C, D) 中。同样可得, 主机 A 备份数据块 (D_1, D_2, D_3), 主机 B 备份数据块 (D_1, D_3, D_4), 主机 C 备份数据块 (D_1, D_2, D_4), 主机 D 备份数据块 (D_2, D_3, D_4)。

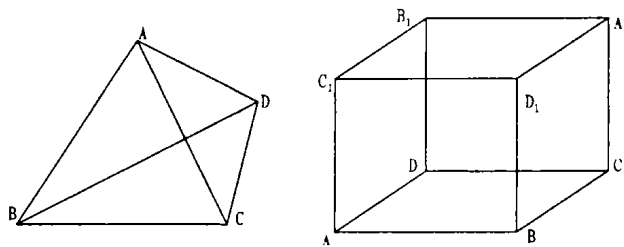


图6 四面体存储模型

图7 六面体存储模型

根据不可恢复闭包定义, 可得不可恢复闭包集合 U 为 $\{\{ABC\}, \{ACD\}, \{ABD\}, \{BCD\}\}$; 根据可恢复闭包定义, 可得可恢复闭包集合 C 为 $\{\{AB\}, \{AC\}, \{AD\}, \{BC\}, \{BD\}, \{CD\}\}$ 。

为了分析结果, 假设每台主机出现故障的概率均为 p , 依据多维算法, 可得:

$R = 1 - 4p^3 + 3p^4, P = 4, D = 2/4, S = 3$

如采用二维算法, 在同样的主机数情况下, 将文件分解为2份 (D_1, D_2), 将 D_1 备份到主机 (A, B), D_2 备份到主机 (C, D), 设每台主机故障率为 p , 可得:

$R = (1 - p^2)^2 = 1 - 2p^2 + p^4, P = 2, D = 2/4, S = 2$

如采用一维算法, 在同样的主机数情况下, 仍然假设每台主机故障率为 p , 可得:

$R = (1 - p)^4, p = 4, D = 4/4, S = 1$

3.2 六面体

如图7所示的六面体 $ABCD-A_1B_1C_1D_1$ 。文件被分解为6份, 其中数据块 D_1 备份到主机 (A, B, C, D), 数据块 D_2 备份到主机 (A, B, D_1, C_1), 数据块 D_3 备份到主机 (A, D, B_1, C_1), 数据块 D_4 备份到主机 (A_1, D_1, B, C), 数据块 D_5 备份到主机 (A_1, B_1, D, C), 数据块 D_6 备份到主机 (A, B, C_1, D_1); 同样可得, 主机 A 备份数据块 (D_1, D_2, D_3), 主机 B 备份数据块 (D_1, D_2, D_4), 主机 C 备份数据块 (D_1, D_4, D_5), 主机 D 备份数据块 (D_1, D_3, D_5), 主机 A_1 备份数据块 (D_4, D_5, D_6), 主机 B_1 备份数据块 (D_3, D_5, D_6), 主机 C_1 备份数据块 (D_2, D_3, D_6), 主机 D_1 备份数据块 (D_2, D_4, D_6)。

根据不可恢复闭包定义, 可得不可恢复闭包集合 U 为 $\{\{ABCD\}, \{ABD_1C_1\}, \{ADB_1C_1\}, \{A_1D_1BC\}, \{A_1B_1DC\}, \{ABC_1D_1\}\}$; 根据可恢复闭包定义, 可得可恢复闭包集合 C 为 $\{\{AA_1\}, \{BB_1\}, \{CC_1\}, \{DD_1\}, \{A_1BC_1\}, \{AB_1C\}, \{B_1CD_1\}, \{BC_1D\}, \{B_1AD_1\}, \{BA_1D\}, \{A_1DC_1\}, \{A_1DC_1\}\}$ 。

为了分析结果, 假设每台主机出现故障的概率均为 p , 可得:

$R = 1 - 6p^4 + 12p^5 - 8p^6 + p^8, P = 6, D = (8/3)/8, S = 3$

采用二维算法, 在同样的主机数情况下; 将文件分为2份, 记为 (D_1, D_2), 将 D_1 备份到主机 (A, B, C, D), D_2 备份到主机 (A_1, B_1, C_1, D_1), 假设每台主机故障率为 p , 可得:

$R = (1 - p^4)^2 = 1 - 2p^4 + p^8, P = 2, D = 2/8, S = 4$

在同样的主机数情况下, 将文件分为4份 (D_1, D_2, D_3, D_4), 将 D_1 备份到主机 (A, A_1), D_2 备份到主机 (B, B_1), D_3 备份到主机 (C, C_1), D_4 备份到主机 (D, D_1), 假设每台主机故障率为 p , 可得:

$R = (1 - p^2)^4, P = 4, D = 4/8, S = 2$

采用一维算法, 在同样的主机数情况下, 仍然假设每台主机故障率为 p , 可得:

$R = (1 - p)^8, P = 8, D = 8/8, S = 1$

表1列出了四面体和六面体情况下各种存储模型的性能参数。从表1可以看出,当主机数为4,多维存储模型能获得高

性能;当主机数为8,文件被分为两份的二维算法性能最高。

表1 各种存储模型的性能比较

	四面体			六面体			
	一维算法	二维算法	多维算法	一维算法	二维算法	多维算法	
备份可靠性(R)	$(1-p)^4$	$1-2p^2+p^4$	$1-4p^3+3p^4$	$(1-p)^8$	$1-2p^4+p^8$	$(1-p^2)^4$	$1-6p^4+12p^6-8p^7+p^8$
节点依赖性(D)	1	1/2	1/2	1	1/4	1/2	1/3
文件分解度(P)	4	2	4	8	2	4	6
空间冗余度(S)	1	2	3	1	4	2	3

结束语 本文提出了一种有效的寻找最优化存储策略的方法,根据分布式备份的基本需求,定义了描述条块化策略好坏的性能参数,定量给出了各种存储模型下的计算方法,从空间几何的角度用多面体形象地描述各种存储模型,并研究了四面体、六面体存储模型的各种性能参数和计算方法。在实际应用中,可以根据该方法和实际情况得到相应的优化策略。

参考文献

- 1 Cabrera L-F, Long D D E. Swift: Using distributed disk striping to provide high I/O data rate. Computing System, 1991, 4(4): 405~436

- 2 Atchley S, Soltesz S, Plank J S, Beck M, Moore T. Fault-Tolerance in the Network Storage Stack. In: Proc. of the FAST 2002 Conf. Monterey, California, USA, Jan. 2002, 3~5
- 3 Long D D E, Montague B R, Cabrera L-F. Swift/RAID: a distributed RAID system. Computing Systems, 1994, 7(3): 333~359
- 4 Cao P, Lim S B, Venkataraman S, Wilkes J. The TickerTAIP parallel RAID architecture. ACM Transactions on Computer Systems, 1994, 12(3): 236~269
- 5 Brinkmann A, Salzweid K, Scheideler C. Efficient Distributed Data Placement Strategies for Storage Area Networks. In: Proc. 12 ACM Symposium on Parallel Algorithms and Architectures (SPAA), 2000

(上接第158页)

的类别涉及植物、动物及日常用品。考虑到图像中对象的光照、平移、旋转发生变化的情况,进行了实验。结果表明,该方法对于对象的光照变化和几何变化(尺寸、平移、旋转等)有很

强的鲁棒性。如下图3、4、5所示,查询结果1是没有发生变化的,而查询结果2是发生了几何变化的。实验结果表明,对象查询的效率总体上与对象的旋转角度成反比。旋转角度越小,查询的准确度越高(见图6)。

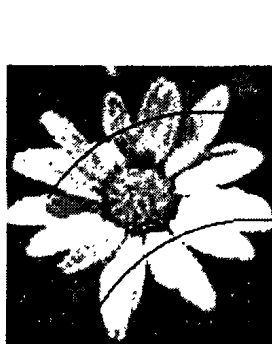


图3 查询对象



图4 查询结果1



图5 查询结果2

表1 查询对象的能力

图像分类	最高	最低	平均
植物	0.941	0.178	0.710
日常用品	0.927	0.182	0.682
动物	0.911	0.142	0.669

性。当然此方法还需进一步改进,如:在选取彩色特征点时,窗口大小的选取;几何约束中角度的选取等,这些是下一步要做的工作。

参考文献

- 1 Rui Y, Huang T S, Chang S-F. Image retrieval: Past present and future. Journal of Visual Communication and Image Representation, 1999, 10: 1~23
- 2 Malki J, Boujemaa N, Nastar C, Winter A. Region queries without segmentation for image retrieval by content. In Visual Information and Information Systems, 1999, 115~122
- 3 Gouet V, Boujemaa N. Object-based queries using color points of interest. In: IEEE Workshop on Content-Based Access of Image and Video Libraries (CBAIVL/CVPR 2001), Kauai, Hawaii, USA, 2001, 30~36
- 4 Montesinos P, Gouet V, Deriche R, Pelé D. Matching color uncalibrated images using differential invariants. Revue Image and Vision Computing, 2000, 18(9): 659~671
- 5 Schmid C, Mohr R. Local gray value invariants for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997, 19(5): 530~534
- 6 Samet H. Object-based and Image-based Image Representations, CMSC 420 Oct. 1998
- 7 魏海, 沈兰荪. 小波变换域内基于方向梯度相角直方图的图像检索算法. 电路与系统学报, 2001, 6(2)
- 8 章毓晋. 图像工程 上册: 图像处理和分析. 北京: 清华大学出版社, 1999
- 9 贾云得. 机器视觉. 北京: 科学出版社, 2000

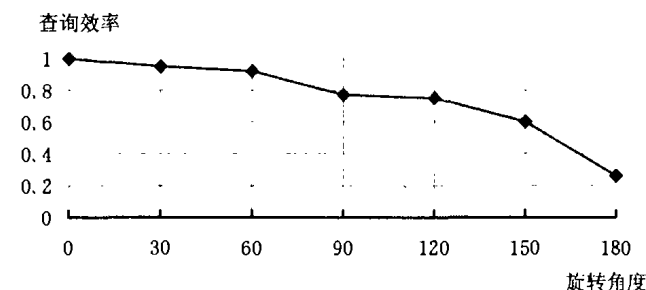


图6 对象的旋转角度及查询效率

总结 本文针对目前的对象查询的发展情况,提出了利用彩色特征点进行查询的方法。不仅克服了分割图像的局限性,而且不以灰度图像为基础进行处理。所采用的彩色特征点及几何约束,对于对象的光照变化和几何变化有很强的鲁棒