Threshold coterie system and its applications

Zhang wei, Liu shuguang

Engineering Institute of the Armed Police, Xi'an 710086, China

Abstract: Give the concept of m-coterie and a method to construct m-coteries. The construction is based on block design and cyclic different set. Application of m-coteries in distributed storage systems is also discussed. Provide a read/write protocol of erasure coding storage systems that use m-coteries to offer redundancy.

Key words: quorum system; coterie; survivable storage

门限 Coterie 系统的构造及应用

张薇，柳曙光

武警学院 电子技术系，陕西 西安 710086

【摘要】针对基于纠删编码的分布存储系统，利用循环差集构造了一类门限 coterie 系统，该系统可以用于构造存储冗余机制，设计了基于门限 coterie 系统的数据读写协议。

【关键词】分布存储系统；容错；门限 Coterie 系统；循环差集

1 引言

Quorum 系统是一个全集上某些子集的集合，这些子集之间满足交汇性（intersection property），即其中任意两个子集的交集不为空，此时每个子集称为一个 quorum. Quorum 系统利用冗余设计来提供容错性能，基于 quorum 的容错技术具有高可用性、并行性、灵活升级等优点，因此得到了广泛的研究和应用.

在 quorum 系统中，要求任意子集间满足交汇性，这样会引来大量的重复和冗余，因为这些子集之间可能存在包含关系，为减少冗余，可以定义一个“精简”的 quorum 系统，其中每个 quorum 都不是其它 quorum 的真子集，这种系统称为 coterie.

Frolund 和 Merchant 等提出了门限 quorum 系统（*m*-quorum system）[1]的概念，其中任意两个子集的交集中至少包含 *m* 个元素，这种系统可以用在基于纠删编码的存储系统中，然而文献[1]中并没有给出 m-quorum 的具体构造方法。

本文利用区组设计的思想来研究门限 quorum 的构造问题，提出了基于循环差集的构造方法，并探讨了门限 coterie 在分布存储系统中的应用。

2 门限 Coterie 系统及其构造

基金项目：国家自然基金 60842006.

2.1 门限 Coterie 系统的定义

定义 1[2] 令 U 为网络系统中全体节点的集合，集合系统 Q 构成一个 Coterie 当且仅当以下条件满足：

(1) 对任意 G∈U，G∈Q 蕴含着 G 非空；

(2) 对任意 G, H∈Q，G∩H≠Φ；

(3) 不存在 G, H∈Q，使得 $G \subset H$.

结合文献[1]中的 m-quorum，我们给出 m-coterie 的定义如下：

定义 2 令 U 为网络系统中全体节点的集合，m 为正整数，集合系统 S 构成一个 m-coterie 当且仅当

(1) 对任意 G∈U，G∈Q 蕴含着 G 非空；

(2) 对任意 G, H∈Q，|G∩H|≥m；

(3) 不存在 G, H∈Q，使得 $G \subset H$.

根据定义，可以给出门限 coterie 的直接构造方法如下：

设 n 为结点个数，m 为门限值，令

$$Q = \left\{ G : G \subseteq U \ and \ |G| \geq \frac{n}{2} + m \right\}$$

显然 Q 是一个门限 coterie，这种构造方法虽然简便直观，但引入了过多的冗余，降低系统效率。以下我们基于区组设计的思想给出一种新的构造方法。

2.2 基于循环差集的构造

定义 3　区组设计[3]

设 S={1,2,…,v}为包括 v 个不同元素的基集，$B=\{B_1,B_2,\cdots,B_k\}$为 S 的 k－子集的集合，r 为含有某一任意元素的 k－子集数，则(S,B)构成一个区组设计，且如果任意一对区组 i, j(i,j=1,2,…,b, i≠j)，有 λ_b 个元素相同，则称为异元均衡区组设计（Differential Balanced Block Design），记作 DBBD(v,b,r,k,λ_b)。

由门限 coterie 的定义，不难看出，DBBD(v,b,r,k,λ_b)与门限 coterie 系统在结构上有较多共性，事实上，DBBD(v,b,r,k,λ_b)相当于一个 λ_b-coterie 系统。因此，DBBD 的构造方法也可以用于构造门限 quorum 或门限 coterie 系统。

定义 4 循环差集（cyclic difference set）

设

$$D\equiv\{a_1,a_2,\cdots,a_k\}\bmod n$$

为以正整数 n 为模的 k 个不同余的整数所组成的集合，如果对任意 $d\neq 0 \mod n$，在 D 中恰好有 m 个有序对 (a_i,a_j)，使得

$$d\equiv a_i-a_j \mod n$$

则称 D 为一个 (n,k,m)循环差集，记作 $D(n,k,m)$.

定理 1　设 n 为节点个数，$D=\{a_1,a_2,\cdots,a_k\} \mod n$ 为 (n,k,m) 循环差集，对任意 i ∈ {1,2,…,n-1}，定义 $B_i=\{a_1+i,a_2+i,\cdots,a_k+i\} \mod n$，则 $B=\{B_1,B_2,\cdots,B_n\}$构成一个 m-coterie.

证明：由循环差集的性质，对任意 B_i, $B_j\in B$, 在 D 中存在 m 对 (a_u,a_v)，u,v∈{1, 2, …, k}，使得

$$j-i\equiv a_u-a_v \bmod n$$

即

$$a_u+i\equiv a_v+j \bmod n,$$

而 $a_u+i\in B_i$，$a_v+j\in B_j$，故 $|B_i\cap B_j|=m$.

例如，设 n=11，D={2,6,7,8,10,11}是一个(n,k,m)=(11,6,3)循环差集，构造各个子集如下：

B1={2,6,7,8,10,11}

B2={3,7,8,9,11,1}

B3={4,8,9,10,1,2}

B4={5,9,10,11,2,3}

B5={6,10,11,1,3,4}

B6={7,11,1,2,4,5}

B7={8,1,2,3,5,6}

B8={9,2,3,4,6,7}

B9={10,3,4,5,7,8}

B10={11,4,5,6,8,9}

B11={1,5,6,7,9,10}

集合{B_1, B_2, …, B_{11}}便可构成一个 3-coterie.

与直接构造相比，基于循环差集的构造方法可以大大提高系统的效率。

3　门限 Coterie 的应用

Quorum 系统通常用于构造互斥系统和存储管理系统. 在分布存储系统中，为提高数据服务的可靠性，常采用备份、纠删编码与门限方案等方法处理数据. 与备份相比，纠删编码和门限方案在安全性和效率上具有较大优势. 纠删编码可以提供数据冗余而避免复制所带来的系统负载. 门限方案则可以避免多个副本的存在带来的安全隐患. 出于对安全性与效率的考虑，近年来人们更倾向于用纠删编码和门限方案取代备份来构造存储系统. 然而这两种方案的使用却带来了新的可靠性问题.

在基于纠删编码和门限方案技术的存储系统中，只有当未失效服务器数量大于某个门限值 m 时，才有可能恢复数据. 用户向 quorum Q_1 中写入数据，从 quorum Q_2 中读出，为了正确恢复数据，Q_1 与 Q_2 的交集中至少要包含 m 个服务器. 针对这种情形，这样可以保证任意选择的读写 quorum 中有足够多的交集来恢复数据。[15]中提出了门限 Byzantine Quorum 系统，在保证读写 quorum 中有足够多交集的同时，还能在部分服务器发生 Byzantine 故障时保持数据服务的持续性。

在基于纠删编码和门限秘密共享的分布存储系统中，门限 coterie 可以用于构造冗余机制，使得任意两个 coterie 的交集中均含有多于门限值的结点，从而实现容错。我们利用 m-coterie 系统来构造数据冗余，在

978-1-935068-01-3 © 2009 SciRes.

写数据时，选择子集 B1，向其中的每个存储结点写入一个数据份额；读数据时，再选择子集 B2，从其中每个结点读取数据并运行恢复算法来得到原始数据。根据门限 coterie 的性质，$|B_1 \cap B_2| = m$，因此可以确保数据的正常恢复。

数据的读写协议如下：

设系统中保存的数据变量为 x，每个客户有一个时间戳 τ，每个服务器 P_i 上则保存着本地副本 x_i 和本地时间 τ_i。

——客户要将数据 x 写入系统时，挑选一个 Quorum Q，将 τ 值加 1，发送（write, x, τ）到 Q 中所有的服务器；服务器接收到消息之后，如果 $\tau > \tau_i$，则更新数据 $(x_i, \tau_i) \leftarrow (x, \tau)$，返回 ACK 信息；客户等待 Q 中所有服务器传来应答信息后结束写操作。

——读操作，客户选择 Quorum Q，向其中所有服务器发送读请求，服务器收到后返回 (x_i, τ_i)，客户收到所有服务器的返回信息后，选择其中具有最大时戳者作为最终读取的数据。

4 结论

本文提出了门限 coterie 系统的定义、构造方法及应用，m-coterie 系统本质上是一种增强了交汇性的 quorum 系统。利用区组设计的思想给出了 m-coterie 的一种构造方法，讨论了 m-coterie 在分布存储系统中的应用。

References(参考文献)

[1] S Frolund, A Merchant, U Saito, S Spence, A Veitch. A decentralized algorithm for erasure-coded virtual disks[R]. Technical Report HPL-2004-46, HP Labs, 2004.

[2] Garcia-Molina H and Barbara D. How to assign votes in a distributed system[J]. Journal of the ACM, Vol. 32, No. 4, pp. 841-860, 1985.

[3] 靳蕃，陈志. 组合编码原理及应用[M]. 上海科学技术出版社，1995.

[4] 张薇，马建峰，王良民，郭渊博. 门限 Byzantine Quorum 系统及其在分布式存储中的应用[J]. 电子学报，Vol.36，No.2，2008.

[5] Jay Wylie, Michael Bigrigg, John Strunk, Gregory Ganger, Han Kiliccote, Pradeep Khosla. Survivable Information Storage Systems[J]. IEEE Conputer, August 2000. 33(8):61-68.

[6] Zhang Wei, Ma Jianfeng. A Reliable and Asymmetric Data Distribution Scheme for Survivable Storage[J]. China Communications, 2006, 3(4): 70-75.