基于兴趣划分的内容发布订阅系统关键算法

逯 鹏 刘旭东 林学练 王 斌

(北京航空航天大学 计算机学院, 北京 100083)

摘 要: 在基于内容发布订阅系统中,将订阅兴趣在多个代理之间划分是代理负载均衡的重要方法. 提出了基于事件空间的 K-D 树划分方法. 通过将事件空间划分成负载相同的区域,优化了系统负载均衡的性能. 基于逻辑空间最短距离的概念提出了与划分相关的事件和兴趣路由算法以及单播和多播混和的通知路由方法. 两种路由算法减少了事件匹配操作,提高了路由的效率. 利用事件空间的区域合并和分裂实现了系统的自组织. 实验和与相关工作比较表明,划分方法及其相关算法的引入提高了系统的可伸缩、容错和负载均衡性能.

关 键 词: 分布式系统; 路由器; 负载均衡; 网络协议

中图分类号: TP 393.03

文献标识码: A **文 章 编 号**: 1001-5965(2006)08-0992-06

Key algorithm in content-based publish/subscribe system based on subscription partitioning

Lu Peng Liu Xudong Lin Xuelian Wang Bin

(School of Computer Science and Technology, Beijing University of Aeronautics and Astronautics, Beijing 100083, China)

Abstract: Partitioning subscriptions interest among multi-brokers acts as an important way to resolve load balancing problem in content-based publish/subscribe system. A new method of event space-based subscription partition with K-D tree was proposed. With this method, the event space was partitioned into zones with equal load and so the system performance of load balancing was improved. Based on the minimum distance of logical event space, new routing algorithms were proposed for event and subscription routing and another hybrid of unicast and multicast routing policy for notification routing. All these algorithms and policy significantly reduced the event matching cost and promoted the routing efficiency in content-based publish/subscribe system. At the same time, the method of splitting and merging zones of event space was used to realize the self-organizing of the publish/subscribe system. The experiment and related works show that the introduction of partitioning methods and related algorithms can improve the scalability, fault tolerant and load balancing performance of content-based publish/subscribe system.

Key words: distributed computer systems; routers; load balancing; network protocols

发布订阅(PS,Publish/Subscribe)系统反映了由事件驱动的信息交互应用的本质特征[1],成为学术和工业界当前研究的热点[2-6],目前分为 2 种类型[1]:基于主题和基于内容. 基于内容系统更通用和灵活,成为研究的重点. 典型系统包括 Elvin[2], Gryphon[3],Siena[4]和 JEDI[5]等. 研究中存在 3 个关键问题[6]:事件匹配,路由,代理网络自组织和负载均衡.

本文主要研究如何基于兴趣划分的方法简化事件路由和匹配操作,并实现系统的自组织和负载均衡. 具体包括:基于 K-D 树的事件空间划分方法,基于兴趣代理的事件和兴趣路由算法,单播和多播混合的通知路由策略,以及系统的自组织协议.

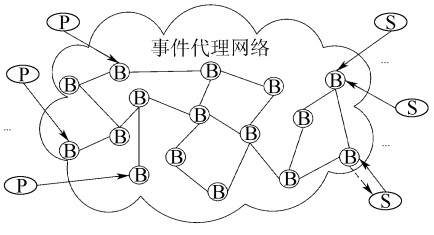
收稿日期: 2005-10-10

基金项目: 国家自然科学基金资助项目(90412011); 国家 863 计划资助项目(2003AA119030)

作者简介: 逯 鹏(1974-), 男, 河南郑州人, 博士生, lupeng@act.buaa.edu.cn.

1　相关工作分析

典型的 PS 系统如图 1 所示. 文献[1]介绍了该过程的相关研究问题.



⟶ 发布 / 订阅　⇢ 通知　P: 发布者　S: 订阅者　B: 事件代理

图 1　广域多代理网络 PS 系统结构

目前兴趣划分的主要方法——基于过滤器划分(FBP, Filter-Based Partition)易造成重复事件流、负载不均衡以及弱可伸缩性等问题,Sinea 是代表系统. 文献[7]提出了基于事件空间划分(ESP, Event Space-based Partition)的方法,其主要动机是:可控的兴趣分布可以实现负载均衡和避免冗余事件流,但文献[7]中的均匀网格划分方法并不能实现对数据非均匀分布应用系统的负载均衡. 本文基于 K-D 树方法解决了该问题,并实现了相关的路由和系统自组织算法.

传统 PS 路由采用泛洪和匹配优先法[8]. 对于平均路由路径 m 较大的系统,2 种方法的效率较低. 本文采用了单播和多播混和的方法,简化了路由算法并提高了路由效率. 此外,典型系统多数是静态配置[6]的,文献[9]虽然给出了基于兴趣划分的 PS 系统解决方案,但其拓扑结构限制为静态配置的层次型. 本文基于 K-D 树的动态自组织协议,实现了网状系统的动态配置.

2　事件空间划分及委派

2.1　符号定义

令事件内容的 N 个属性分别代表事件空间 Ω 中的一维,则单个事件 $e \in \Omega$ 是 Ω 中的一个点. 设兴趣是由不同维的半开半闭区间(例如 $100 < X_{\text{price}} \leqslant 150$)构成的超矩形区域 $z \subseteq \Omega$. 所有区域构成集合 Z,系统中代理的集合为 B, $|B| = K$;区域 $z_i \in Z(1 \leqslant i \leqslant K)$ 的负载是 $M(z)$.

2.2　事件空间划分

K-D 树划分的算法思想是:首先沿空间 Ω 的 $d_t(1 \leqslant t \leqslant N)$ 维将 Ω 划分成 2 个负载相同的区域 $M(z_1) = M(z_2)$;二叉树的 2 个叶子节点指向区域 z_1, z_2,父节点指向 2 个孩子节点;然后,按照层次遍历的方法递归此过程,每个层次依次沿 Ω 的不同维划分区域并循环,直至形成 K 个叶子节点,如算法 1 所示. 其中负载函数 M 与应用领域事件的概率分布相关,例如文献[7]中的类 zipf 分布.

算法 1　事件空间划分算法

输入:事件空间 Ω,代理数量 K

输出:空间划分结果队列 Q

步骤:

1) Initiate(Q);//初始化一个队列
2) root = Ω;
3) Enqueue(Q, root);
4) $d = 1$;
5) While (Q. length < K) {
6)　P = Dequeue(Q);
7)　P = SplitByLoad(P, d);
8)　Enqueue(P. leftchild);
9)　Enqueue(P. rightchild);
10)　$d = \lfloor \log_2 Q.\ \text{length} \rfloor \bmod N$;
11) }
12) return Q

算法 1 首先将 Ω 作为根节点(第 1) ~ 3)行),然后判断队列中的节点是否小于 K(第 5)行),是则将当前节点划分为 2 个负载相等的节点,当前节点成为其父节点(第 6) ~ 9)行);如果本层节点完成划分则依次沿不同维循环划分(第 10)行). 算法完成,队列 Q 中包含了 K 个负载相等的区域 z_i,如图 2 所示. 图 3 给出了相应的划分树,其中 $\Omega = \bigcup_{i=1}^{k} z_i$, $z_i \cap z_j = \varnothing$, $1 \leqslant i, j \leqslant K$. 区域的坐标由维上的区间构成,如图 2,设矩形区域 z_4 的坐标为(2 ~ 6, 4 ~ 5).

集合 Z 中的元素有以下关系:

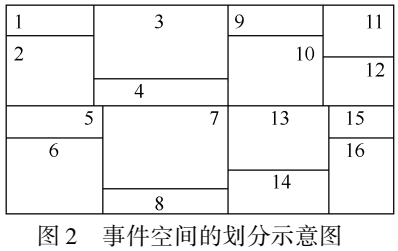


图 2　事件空间的划分示意图

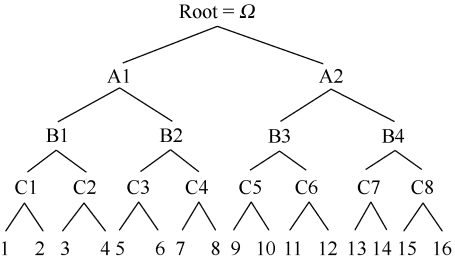


图 3　事件空间 K-D 树划分逻辑结构

定义1 区域相邻关系

设 $z_1, z_2 \in Z$,如果 z_1, z_2 在任意一维上坐标相邻,在其余维上相交,则称 z_1, z_2 区域相邻,记作 $z_1 R z_2$.

如图2,$z_5 R z_6$,$z_2 R z_4$,但 z_4 与 z_{13} 不相邻.

定义2 兄弟区域

设 $z_1, z_2 \in Z$,如果 z_1, z_2 在任意一维上坐标相邻,在其余维上相同,则称 z_1, z_2 互为兄弟区域,记作 $z_1 S z_2$.

如图2,$z_1 S z_2$,$z_2 R z_3$ 但不是兄弟区域.

定理1 $z_1, z_2 \in Z$,如果 $z_1 S z_2$,那么 $z_1 R z_2$.

证明略.

定理2 算法1形成的K-D索引树中2个兄弟叶子节点指向的区域是兄弟区域.

证明

因为2个叶子节点指向的区域 z_1, z_2 是同一区域沿空间中任意一维划分的结果,由定义2,$z_1 S z_2$. **证毕**

以下用节点代表其所指向的区域.

定理3 $x \in Z \rightarrow (\exists y, u \in Z)(xSy \vee (xRy \wedge ySu))$.

证明

由二叉树性质,任意节点 $x \in Z$,其相邻节点是叶子节点或者中间节点.

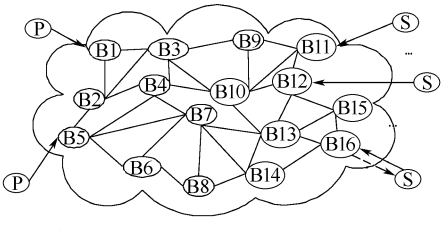
如果是叶子节点 y,由定理2,即 $x \in Z \rightarrow (\exists y)(xSy)$.

如果是中间节点 v,则由算法1中的二叉树层次划分性质,节点 v 一定存在2个孩子节点即 $(\exists y, u \in Z)((z = y \cup u) \wedge (ySu))$. 因为 $v = y \cup u$ 所以 $xRv \rightarrow xRy \vee xRu$,不妨设 xRy,则 $x \in Z \rightarrow (\exists y, u \in Z)(xRy \wedge y\delta u)$ 成立. **证毕**

定理3表明事件空间中的任意区域都具有兄弟区域,或者有一个相邻区域具有兄弟区域.

2.3 兴趣区域委派

兴趣区域委派定义为映射 $f: Z \rightarrow B$. f 将区域分别委派系统中的代理,并将区域的相邻关系映射为代理的链接关系. 如图4是图2的映射结果. 称每个区域的代理为兴趣代理,兄弟区域的代理为兄弟代理,相邻区域的代理为相邻代理. 代理之间的逻辑距离即区域中心 p_1, p_2 间的距离为 dstnt.



→ 发布/订阅 --→ 通知 P: 发布者 S: 订阅者 B: 事件代理

图4 基于事件空间划分的事件代理网络

3 路 由

每个代理的邻接路由表如表1所示. 系统路由包括事件路由、兴趣路由和通知路由.

表1 兴趣代理 B7 的区域路由表

相邻区域	相邻代理
z_4	B4
z_5	B5
z_6	B6
z_8	B8
z_{13}	B13
z_{14}	B14

3.1 事件和兴趣路由

事件和兴趣路由是发布者和订阅者将事件和兴趣路由到兴趣代理的过程. 其基本算法思想是:基于事件空间,选择与目标代理逻辑距离 dstnt 最短的相邻代理作为下一跳地址,如算法2所示.

算法2 事件路由算法

输入:事件 e

输出:无

步骤:

1) if e. source = Local _ client {
2) e. target = Lookup _ Match _ Zones(e)
3) }
4) forward(e);

算法2首先计算匹配 e 的兴趣区域(算法第1)~3)行),然后调用 forward 函数转发事件 e. 如果事件 e 来自相邻代理则直接转发(第4)行). 算法2完成了 e 的区域匹配和转发.

算法3 兴趣路由算法

输入:订阅兴趣 sub

输出:无

步骤:

```
1)  if sub.source = Local_client {
2)      zones = Lookup_Overlap_Zones(sub.zone);
3)      for each z in zones{
4)          b = broker_for_zone(z);
5)          newsub = intersection(z, sub.zone);
6)          if(newsub ≠ null){
7)              newsub.targetbroker = b;
8)              forward(newsub)
9)          }
10)     }
11) }else{
```

```
12)    forward(sub)
13) }
```

由于 sub 与划分区域不完全重合. 因此,算法 3 首先计算与 sub 相交的区域(第 1),2)行),并将 sub 分解为一组兴趣 newsub,然后分别转发(第 3) ~10)行);若 sub 源于相邻代理则直接转发(第 11),13)行). 算法 3 完成了兴趣 sub 的区域匹配和转发.

算法 4　forward 函数算法

输入:事件 e(或者兴趣 sub)

输出:无

步骤:

```
1)  dstnt = -1;
2)  nexthop = Local;
3)  if e. target = Local then{
4)    receive(e);
5)  }else{
6)    for each broker in neighborlist{
7)      temp = distant(broker, e. target);
8)      if dstnt<temp {
9)        dstnt = temp;
10)       nexthop = current_broker;
11)     }
12)   }
13)   if dstnt>0 {
14)     send e to nexthop;
15)   }
16)}
```

算法 4 首先判断事件 e(或者兴趣 sub)的目标代理是否为本地(第 1) ~5)行),是则接收;否则,查找并计算相邻代理中距目标代理最近的代理(第 5) ~12)行),然后转发(第 13) ~15)行). 算法 4 完成了查找下一跳地址的操作.

以图 4 为例,当 B5 接收到 e,它首先匹配 e 的兴趣代理是 B11;第 2 步,B5 计算距离 B11 最近相邻代理是 B4,转发 e 到 B4;第 3 步,B4 接收到 e,判断 B10 是最短距离代理,转发 e 到 B10;第 4 步,B10 判断 B11 是其邻居转发 e 到 B11,完成事件路由.

3.2　通知路由

通知路由是 Bi 根据本地兴趣集合 Bi. subs,为 $e \in z_i$ 细粒度匹配订阅者,并通知所有订阅者的过程.

事件匹配通常采用兴趣索引树的方法,匹配过程是 e 从根开始搜索匹配树的过程. 基于 ESP,z_i 将需要匹配的兴趣数量平均缩小为总量的$1/K$,提高了事件匹配效率和通知路由的效率.

通知路由分为单播和多播 2 种情况. 设置参数 t,例如 $t = \bigcup_{i=1}^{k} |\mathrm{B}i.\ \mathrm{subs}|/K$,即订阅者数量的平均值. $|\mathrm{B}i.\ \mathrm{subs}| < t$ 则采用算法 2 单播;$|\mathrm{B}i.\ \mathrm{subs}| > t$,则建立应用层多播树,使用多播函数 multicast(e)多播. 设置参数 t 减少了建立多播树的系统开销,具体的多播算法作为后续工作讨论.

4　自组织策略

本节主要提出系统自组织相关的可伸缩协议和容错协议. 可伸缩协议主要解决兴趣代理的加入和退出时系统的自组织问题;容错协议主要解决链路和兴趣代理失效时系统的自愈和恢复问题.

4.1　可伸缩协议

4.1.1　代理的加入

从事件空间的角度,新代理加入可以理解为一个区域分裂的过程,如图 5 所示. 首先,假定新代理 By 能够发现系统中的代理 Bx,则系统加入 By 的过程如下:

1) By 向 Bx 发送加入消息;

2) Bx 将区域分裂为 z_x,z_y,并将 z_y 及相关订阅者列表委派 By;

3) Bx,By 分别重新计算相邻代理,并通知相邻代理更新路由.

以图 5 为例,假设加入代理 B17 建立了与 B13 的连接. 首先 B17 向 B13 发出加入的消息;B13 将其区域分裂为 2 个区域 z_x,z_y(沿该区域最后一次分裂的后继维进行划分),B13 保留区域 z_x,将区域 z_y 委派给 B17,并将相关的订阅者列表复制到 B17;B13 和 B17 重新计算相邻代理(可以通过删除原路由表中失效路由项的方法实现),并通知相邻的代理更新路由信息,完成代理 B17 的加入操作.

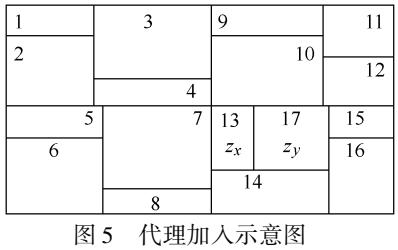


图 5　代理加入示意图

4.1.2　代理的退出

代理的退出可以理解为事件空间中 2 个区域合并的过程. 该过程分为 2 种情况,如果 Bx 具有

兄弟代理 By,则退出步骤如下:

1) By 合并 Bx 的区域并复制其订阅者和路由列表;

2) 删除重复路由项,通知相邻代理更新路由.

如图5所示,如果 B7 退出,它首先判断 B8 是它的兄弟区域,B7 向 B8 发出退出消息,其中包含了订阅者的列表和路由表;B8 更改其区域和路由表,并通知相邻的代理更新路由,完成 B7 的退出操作.

否则采用"替换"方法,基本策略是:由定理3,Bx 一定存在相邻代理 By,且 By 具有兄弟区域 Bu,称 By 为替换代理;将 By 与 Bu 合并,由 Bu 代理合并后的区域;By 则替换 Bx,Bx 退出系统. 具体步骤如下:

1) Bx 向相邻代理发送退出消息,得到 By 响应;

2) Bx 向 By 发出退出请求;

3) By 与兄弟代理 Bu 进行合并,合并后的区域委派 Bu;

4) By 复制 Bx,此时 By 的角色转换为原代理 Bx 的角色,包括代理标识;

5) 原 Bx 退出系统.

如图5所示,如果 B14 退出,因为兄弟代理的区域已经分裂,所以发出退出消息,获得替换代理 B8 的响应,B14 向 B8 发出退出请求,B8 则首先合并兄弟代理 B7;然后复制 B14,并更新路由信息. 退出只是底层代理的转移,除合并兄弟区域外,并没有改变多维空间区域的相邻关系.

4.2 容错处理

4.2.1 链接失效处理

如果代理间连续 n 次超时重发未得到响应,则认为链接失效. 由于每个代理具有多个链接,因此当检测到链路失效,路由算法自动选择其余代理中距离最近的作为下一跳地址转发事件. 如图4所示,若 B5-B4-B10-B11 的传输路径当中,B5 与 B4 之间的链路失效,则路由算法基于最短距离,自动选择 B2-B3-B9-B11 作为新的传输路径.

4.2.2 代理失效处理

如果代理 By 与相邻代理 Bx 间的所有链接失效,并且 Bx 是事件 e 的目标代理,无法通过链接失效处理,则认为 Bx 失效. 基本的处理策略是:如果 By 是 Bx 的替换代理,则启动替换时钟,时钟到期则通过替换策略合并 Bx,并向相邻的节点发送停止时钟的消息,完成失效处理;如果 By 是 Bx 的兄弟代理,则直接合并 Bx,并向相邻代理发送停止时钟的消息,完成失效处理. 时钟周期设置与代理负载成正比,负载大则周期长,保证较低负载的代理首先进行失效处理. 如图5,当 B7 失效,则 B8 将其区域合并;如果 B14 失效,则 B13 通过替换的策略将 B14 接管.

5 仿真实验和分析

基于 NS2 工具产生图4所示的代理网络,对不同划分策略进行了实验. 函数 $e=g(x,y)$ 作为事件源,x,y 的取值分别符合 zipf 分布. 首先将属性 x,y 的值域所构成的二维空间 Ω 分成16个区域相等的网格,每个网格委派给一个代理. 实验中,将落入每个网格内的 e 发布到相应的代理,测试代理的负载. 然后,利用函数 $e=g(x,y)$ 产生4000个数据,按照 K-D 树的划分方法,将数据划分为数量相等的16个区域,每个区域委派给不同的代理,然后重新产生数据,并将落入相应区域的事件发送其代理.

实验中设置了一个发布端和一个订阅端模拟产生相应事件的发布和订阅者进程. 将已经产生的数据作为事件向目标代理发布. 代理处理数据路由的负载权重为1,处理数据匹配的负载权重为10,即该数据的目的地是该代理. 为了降低随机对实验结果的影响,进行了10次测试,取其平均值作为代理的负载. 2个数据评价指标:最大/最小负载比和最大负载量. 实验结果见图6和图7.

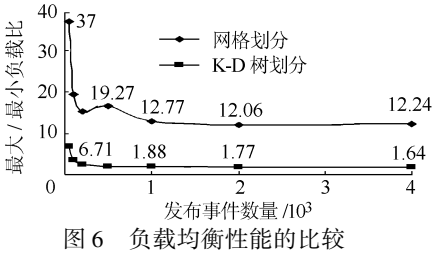


图6 负载均衡性能的比较

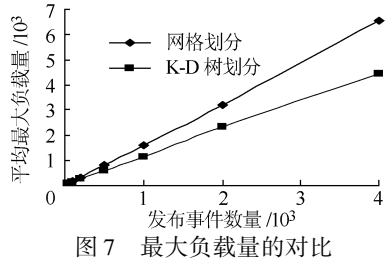


图7 最大负载量的对比

图6可见,随着事件发布数量的增加,K-D 树划分的方法产生的最大/最小负载比趋于1,而网格划分的最大/最小负载比趋于12,两者存在较大的差距. 说明 K-D 树划分的方法有效均衡了事

件的负载. 图 7 可见,两者中代理最大的负载量的增长趋势,K-D 树划分的方法要明显低于网格划分的方法.

6　结 束 语

本文基于 K-D 树解决了事件非均匀分布情况下的兴趣划分问题,均衡了代理的负载,简化了路由,提出了系统的自组织策略和方法. 后续的工作是利用区域的合并和分裂以及代理的加入和退出机制,实现系统的负载均衡机制,并且考虑将技术实现嵌入自主开发的信息交换中间件 ACT-IIP,使其成为 web 服务和网格服务等上层应用的通信基础设施.

参考文献 (References)

[1] Eugster P Th, Felber P A, Guerraoui R, et al. The many faces of publish/subscribe[J]. ACM Comput Surv, 2003, 35(2): 114-131

[2] Fitzpatrick G, Kaplan S, Mansfield T, et al. Supporting public availability and accessibility with Elvin: experiences and reflections[J]. Comput Supported Coop Work CSCW Int J, 2002, 11(3-4): 447-474

[3] IBM Corporation. Achieving scalability and throughput in a publish/subscribe system [R]. RC23103(W0402-026), 2004

[4] Carzaniga A, Rosenblum D S, Wolf A L. Design and evaluation of a wide-area event notification service[J]. ACM Trans Comput Syst, 2001, 19(3): 332-383

[5] Cugola G, Di Nitto E, Fuggetta A. The JEDI event-based infrastructure and its application to the development of the OPSS WFMS[J]. IEEE Trans Software Eng, 2001, 27(9): 827-850

[6] 薛涛,冯博琴. 内容发布订阅系统路由算法和自配置策略研究[J]. 软件学报, 2005, 16(2): 251-259
Xue Tao, Feng Boqin. Research on routing algorithm and self-configuration in content-based publish-subscribe system [J]. Journal of Software, 2005, 16(2): 251-259 (in Chinese)

[7] Riabov A, Liu Zhen, Wolf J L, et al. Clustering algorithms for content-based publication-subscription systems [C]//Proc Int Conf Distrib Comput Syst. Piscataway, NJ: IEEE, 2002: 133-142

[8] Banavar G, Chandra T, Mukherjee B, et al. Efficient multicast protocol for content-based publish-subscribe systems[C]//Proc Int Conf Distrib Comput Syst. Piscataway, NJ: IEEE, 1999: 262-272

[9] Wang Y M, Qiu L, Achlioptas D, et al. Subscription partitioning and routing in content-based publish/subscribe networks[C]// Dahlia M. 16th International Symposium on Distributed Computing. Berlin: Springer-Verlag, 2002: 28-30

(责任编辑:文丽芳)

(上接第 968 页)

如果不采用光褪色加以防护,同样的光纤环,0.2 kGy 的总剂量导致 9.58 dB 的损耗增加量[6],这对于光纤陀螺的系统解调来说是不允许的. 光纤环是光纤陀螺的敏感部件,光纤环损耗的变化,对陀螺性能影响非常大,甚至不可用. 本次实验充分说明光褪色的效果.

由上面的实验和分析知道,利用 LD 光褪色可以改善光纤环及陀螺内所有光电器件的光纤尾纤的抗辐照性能. 但是光纤陀螺内的光为恒定的,加入光会影响陀螺的解调,导致陀螺不可用.

5　结 束 语

本文通过实验,得到了保偏光纤环的辐照效应,验证了利用光褪色可以提高光纤的抗辐照性能,并首次提出将光褪色应用到光纤陀螺中作为其空间应用的抗辐照加固手段,具有非常好的应用前景.

参考文献 (References)

[1] Richard H B, Warren F W, Terrenc S L, et al. Proton-induced degradation in interferometric fiber optic gyroscopes [C]//Taylor E W. Photonics for Space Environments III. Bellingham, WA: SPIE, 1995: 120-163

[2] Skutnik B j, Greenwell R A, Scott D M, et al. Radiation behavior of pure silica core optical fibers[C]//Greenwell R A. Fiber Optics Reliability: Benign and Adverse Environments II. Bellingham, WA: SPIE, 1988: 24-35

[3] Brambani L A, Friebele E J. Radiation effects in polarization-maintaining fibers[C]//Greenwell R A. Fiber Optics Reliability: Benign and Adverse Environments II. Bellingham, WA: SPIE, 1988: 43-49

[4] Gusarov A I, Doyle D B. Modeling of gamma-radiation impact on transmission characteristics of optical glass[C]// Berghmans F. Photonics for Space and Radiation Environments II. Bellingham, WA: SPIE, 2002: 78-85

[5] 李荣玉,殷宗敏. 石英光纤抗辐照加固的研究[J]. 上海交通大学学报, 2000, 34(2): 215-217
Li Rongyu, Yin Zongmin. Research on anti-radiation of silica fiber[J]. Journal of Shanghai Jiaotong University, 2000, 34(2): 215-217 (in Chinese)

[6] 张玉艳,肖文. 卫星用光纤陀螺中抗辐射光纤的研究[J]. 光学与光电技术, 2005, 3(1): 42-45
Zhang Yuyan, Xiao Wen. Research on the radiation-resistant optical fibers for fiber-optic gyroscopes in the satellites[J]. Optics and Optoelectronic Technology, 2005, 3(1): 42-45 (in Chinese)

(责任编辑:张　嵘)