

基于网络阻塞传染路径的关键阻塞结点确定

The Sticking Point Ascertaining of Block Based Infection Route of Network Block

陶 洋

(重庆邮电学院软件中心 重庆 400065)

Abstract Firstly, the appearance of block infection is analyzed in this paper. And then its definitions and description are given. Also, an algorithm of ascertaining the Infected Paths of Blocking is given. Based on them we further put forward the algorithm of sticking point of blocking and analyze the method of breaking off blocking. These algorithms and methods may help us control and break off network blocking better.

Keywords Network, Block infection, Route, Node

一、引言

故障有传染现象^[1],网络阻塞同样有传染现象^[3,8]。网络单元的故障可以引起网络阻塞,而网络的阻塞决不仅由故障引起。由于网络状态与网络上运行的业务及特征有很强的相关性,而业务特性,特别是流量,具有很强的随机性,因此由业务特性变化引起的网络阻塞是一种动态随机现象,具有一定的时效性。为此可将阻塞分为如下两类:

1. 静态阻塞 主要由于网络单元的物理故障或系统软件故障所引起的通信中断或通信能力的下降。一般说来,这类故障具有固定的表现,由此引起的故障特征必然包含该网络网元的特征,且该网元导致了阻塞及扩散现象。

2. 动态阻塞 不同网络业务有不同的特性,特别是业务流速^[4,6],当具有不同流速业务在网络上交织传输时,业务流的流速具有动态随机可变性。而网络各个单元都具有确定的最大业务流的接纳能力,网络阻塞状态又相关于业务特性,因此这类由业务特征的变化而引起的阻塞具有很强的动态时效性及非线性特性,将其归结为动态阻塞。

另一方面,网络可以抽象地描述为由网络结点和网络传输边相互连接而构成,因此网络上的任意一结点的阻塞都将可能引起别的结点或别的结点集合的阻塞,造成系统通信性能的下降,甚至瘫痪。这就是网络阻塞传染现象^[3,8]。

至此,我们可以将引起网络阻塞的原因划分为三个方面:其一,网络结点或网络结点连接的结构、性能、故障等因素而引起的阻塞;其二,路由与流量不协调,包括路由的承载能力和流量分布不平衡等因素;其三,网络阻塞传染,即一个网络结点的阻塞将引起另外一个或一些网络结点的阻塞。

一旦网络发生阻塞现象,首先表现给出的是阻塞路由的一个集合。由于阻塞存在传染的情形,引起阻塞的网络结点或边、被传染的网络结点或边都包含在这个集合之中。因此不论是对阻塞路由的确定还是对阻塞点或边的确定,都必须从阻塞现象的表征分析入手,得到相应的诊断结论,确定网络阻塞及引起传染的关键结点或边。

对于阻塞及阻塞控制的研究人们已经取得了较大的进展,且具有较为丰富的研究成果,而对阻塞传染中阻塞结点确定的相关问题进行研究,到目前为止我们可以借鉴仅有故障传染研究^[1,7],以及我们前期的研究成果^[3,8]。

二、阻塞传染现象及相关定义

按照图论中的等价变换原则,可以将阻塞问题完全归结于网络结点,有利于阻塞传染问题的分析和结论的得出。首先,给出下列一些基本定义:

定义 2.1 给定一个常数 Φ , 在一定时间内结点 i 的资源占用量大于 Φ , 且 $T < 0$, 称该结点为阻塞点。其中结点吞吐率 $T = D_{OUT}/D_{IN}$, T 关于时间 t 的一阶导数 $T' = \partial(D_{OUT}/D_{IN})/\partial t$, 称为该结点的阻塞程度, 简称结点阻塞度, 记为 $B(i)$ 。

定义 2.2 如果一条从起点 S 到终点 T 的有向路径, 包含至少一个阻塞点, 称该路径为阻塞路径。该路径上所有阻塞点阻塞程度之和, 即为该路径的阻塞程度, 简称路径阻塞度, 记为 $\Omega(\lambda)$, 其中 λ 为一条有向路径。

定义 2.3 在一条阻塞路径上, 与终点 T 距离最短的阻塞点为关于该路径的首阻点。

定义 2.4 设有向路径 $(S, \dots, a, b, \dots, l, \dots, T)$, 其中 a, b, l 等为结点, 如果 b, l 为阻塞点, 那么 b 为 l 的后继阻塞点, l 为 b 的前驱阻塞点。

定义 2.5 一个阻塞点属于不同阻塞路径的数量为该点的阻塞关联系数。

阻塞的传染波及整个网络范围, 如图 1 我们将网络表示为一个图 G , 其传输线路的集合为图的边集 E , 交换转接点的集合为图的顶点集 V 。假设 $G_1, G_2, \dots, G_6 \subset G$, 且 $G_1 \cap G_2 \cap \dots \cap G_6 = \Phi$, 并且给出其相关的定义。

定义 2.6 子图 $G_1 \subset G$, 由于 G_1 的阻塞而使得与 G_1 连通的子图 G_2 同样表现出阻塞特征, 称这种现象为阻塞传染现象。

定义 2.7 如果 $G_1(V', E')$ 出现阻塞, 而引起 $G_2(V'', E'')$ 的阻塞, 称 $G_2(V'', E'')$ 为 $G_1(V', E')$ 感染者。

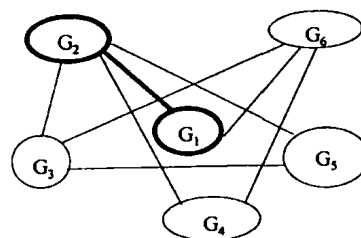


图 1 子图及传染子图

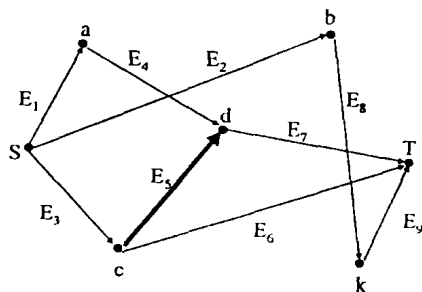


图2 网络流图

对一个网络加载信息流后,将其表示为有向图,如图2。网络的边是无向的,而信息流是有向的。我们描述的是与信息流特征密切相关的阻塞传染现象,故将采用有向图及相关理论进行分析和表达。

如果S向T转移信息,那么在S和T之间可以形成多条有向路径。图2中假设结点d发生阻塞,而结点c与结点d连通,从而使得弧E5信息流通过量减少,且结点c同样表现出阻塞特征,但它却是“被传染的”。

定义 2.8 如果结点i将阻塞现象传染给结点j,称i为传染者;j为被传染者,用(i,j)表示。并且i,j可以是相邻结点,也可以是同一条有向路径上的不相邻结点。

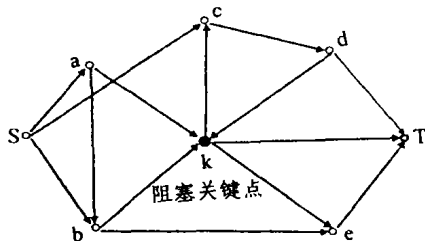


图3 阻塞关键点与阻塞路径

定义 2.9 网络中关联阻塞路径最多的网络阻塞结点,称之为该网络的关键阻塞点。

求出网络上所有阻塞路径所关联的关键阻塞点,即是该网络的关键阻塞点。按确定路由方式,如果它是所有路径的首阻点,它的疏通,即可解除网络的阻塞状态;而随机路由方式下的网络关键阻塞点的求取和网络阻塞的解除更具复杂性。

根据前面的定义可知,任何阻塞结点都必须包含于阻塞路径之中,因此要获得引起阻塞的关键结点方法之一就是先求出所有的阻塞路径。

三、阻塞传染路径算法

通常情况下,图中确定最短路径算法是按路径的增长而弧或边加权值最小取值的原则而设计的,而这类算法不适用于所面临的问题,因为阻塞传染路径的增长不是按最小取值原则进行^[3]。为了解决结点加权情况下的阻塞传染的路径求取,我们分析了包括 Floyd 算法等在内的多种算法,仍不能求得阻塞传染路径^[8]。

由于结点加权求取相应的路径较为困难,因此必须对其相对应的图转换为等价的边或弧加权图,再进行算法设计^[3],得出有用的结论。其相关的约定和算法如下。

定义 3.1 设有向图G不包含回路且其邻接矩阵A元素 a_{ij} 给出如下定义:

1)若结点i与结点j之间存在一条弧,则 a_{ij} 为该弧的阻

塞度;

2)若结点i与结点j之间没有弧存在,则 $a_{ij} = -\infty$ 。

并设路径矩阵的每一元素以路径的集合方式表示。为此我们给出如下求取图中所有阻塞路径的算法A-3.1。

第一步 将图G中的结点进行编号排序。

第二步 按定义3.1,给出图G的邻接矩阵 $A(a_{ij})_{n \times n}$ 。

第三步 路径矩阵 $P^{(0)}(p_{ij})_{n \times n}$ 的初始值为空矩阵,其中 $p_{ij} = \{NULL\}$ 。

第四步 设 $1 \leq k \leq n$,在矩阵 $p_{ij}^{(k-1)}$ 基础上,将编号为k的结点 v_k 插入任意两点 v_i 与 v_j 之间,并按下列原则求出 $p_{ij}^{(k)}$:

(1)结点 v_k 不包含于 $p_{ij}^{(k-1)}$ 中的任何路径。

(2)结点 v_k 的插入不能在 $p_{ij}^{(k)}$ 中的任何路径形成回路,否则在 $p_{ij}^{(k-1)}$ 和 $p_{ij}^{(k)}$ 中的相应路径不做修改。

(3)查询邻接矩阵A,使得 v_k 结点插入 $p_{ij}^{(k-1)}$ 中的路径后形成的 $p_{ij}^{(k)}$ 中阻塞度小于0的路径,仍然是 v_i 与 v_j 之间存在的路径。

第五步 重复第四步直至 $k=n$,即可得到路径矩阵 $P^{(n)}(p_{ij})_{n \times n}$ 。

根据前面的定义,很显然路径矩阵 $P^{(n)}(p_{ij})_{n \times n}$ 包含了图G所有的阻塞路径,元素 p_{ij} 表示结点i(起始点)到结点j(终点)的所有阻塞路径的集合。在此基础上我们就可以通过一定的算法设计求得阻塞的关键结点。

四、关键阻塞结点及路径首阻结点的确定

前面,我们已经求得了包含所有阻塞结点的阻塞传染路径,当然每一条阻塞路径的阻塞程度是不相同的^[3],这是由于每一个阻塞结点的阻塞程度不同而造成的。假如所有的网络连接(边)都不发生阻塞或故障,阻塞结点是引起阻塞的全部原因,而这些结点包含了引起阻塞的关键结点和被传染的阻塞结点,如定义2.9,关键阻塞结点是阻塞路径与之相汇最多的阻塞结点,因此从理论上讲关键阻塞结点的阻塞条件解除,路径的阻塞即解除,但是多条路径的交汇只是引起关键阻塞结点阻塞的主要原因,而关键阻塞结点的前驱阻塞结点也是引起其阻塞的重要因素,我们必须求出关键阻塞结点和其前驱阻塞结点(即首阻点)。

定义 4.1 称如下阻塞状态的表示模型为阻塞症候模型(Blocking Symptom Model,简称BSM模型)。其表示如下:在 Δt 时间内,如 $B(i)$ 表示了结点i的阻塞度,则根据定义2.1有:

$$D(i) = \begin{cases} 1 & i \text{ 为阻塞点。} T' < 0 \\ & \text{且结点资源占用量大于} \Phi。 \\ 0 & i \text{ 正常结点(非阻塞点)。} \end{cases}$$

在讨论关键阻塞结点的确定之前,根据定义4.1所给出的BSM模型,我们先给出如下的定义。

定义 4.2 称矩阵 $F(f_{ij})_{n \times n}$ 为图G的阻塞症候,其中 f_{ij} 为一个二元组 (D_{ij}, B_{ij}) , D_{ij} 表示弧 (v_i, v_j) 的阻塞状态,其值为0或1; B_{ij} 表示弧 (v_i, v_j) 的阻塞程度。

按前面给出的阻塞路径算法,能求得图G中任意两点的阻塞路径,如果结合阻塞症候矩阵 $F(f_{ij})_{n \times n}$ 可以很方便地计算出每一条阻塞路径的阻塞程度值。

在已获得的路径矩阵 $P^{(n)}(p_{ij})_{n \times n}$ 的基础上,取任意两个结点 v_i 和 v_j ,相应的路径集合为 P_{ij} 。

令 $p_{ij} = \{l_{ij}^1, l_{ij}^2, \dots, l_{ij}^m\}$, $m \geq 0$, 其中 l_{ij}^k , $0 \leq k \leq m$ 表示一条

从结点 s 到结点 t 的路径。

再将 L_u 表示为:

$(v_i, v_{i1}, \dots, v_{in}, \dots, v_n)$ 其中 $0 \leq i \leq n-2$

因此对于给定的路径集合 P_u , 我们给出如下算法求出除结点 v_i 和 v_i 之外所有的结点关于该路径集合的阻塞关联系数。

设 Z_i 为包含结点 i 的路径集合, q_i 表示结点 i 关于 P_u 阻塞关联系数。阻塞关联系数算法 A-4.1 如下:

1) 删除 P_u 中各弧阻塞状态之和为 0 的路径, 得阻塞路径集合 P'_u ;

2) 计算 P'_u 中的路径所包含的除结点 v_i 和 v_i 之外所有的结点, 得到集合 N ;

3) 按顺序取 N 中的一结点 i , 计算出相应的 Z_i 和 q_i ;

4) 删除 Z_i 中所有路径的结点 v_i 和结点 v_i ;

5) 重复第 3) 步, 直至取完集合 N 中的全部结点。

算法的结果为集合 N 中的每个结点 i 所关联的路径集合 Z_i 以及关于阻塞路径集合 P'_u 的阻塞关联系数 q_i 。为此可得出如下结论。

定理 4.1 对于任意的 $v_i \in N$, 有 $Z_i \neq \emptyset$ 且 $q_i \geq 1$ 。

因为 N 中的结点是 P'_u 中所有路径上的结点的并集, 但不包含除结点 v_i 和结点 v_i , 因此 N 中的任何一个结点至少属于 P'_u 中的一条路径, 故有 $Z_i \neq \emptyset$ 且 $q_i \geq 1$ 。

对于任意的两条路径 $l_i, l_j \in P'_u, i \neq j$, 相应路径上的结点 (不包含结点 v_i 和结点 v_i) 集合分别为 N_i 和 N_j , 令

$$Y' = \bigcup_{v_i \in N} N_i$$

Y' 可能包含了一些非阻塞结点, 将这些非阻塞点从 Y' 中删除, 得阻塞结点集合 Y 。

定理 4.2 结点集合 Y 为关于阻塞路径集合 P'_u 的关键阻塞点集。

因为任意结点 $u \in Y$, 必然存在一个 N_i , 有 $u \in N_i, u$ 即是路径 l_i 上的一个阻塞点。同理可以证明 u 是 P'_u 中任意路径的阻塞点, 根据定义 2.9, u 是关键阻塞点。定理 4.2 即可得到证明。

我们可以非常容易地将 Y 中的结点按路径距离终点 t 的最小弧数进行排序, 确定出首阻关键结点、次首阻关键结点等。依次对阻塞点的疏通可在一定程度上解除网络的阻塞状态。

关键阻塞点的疏通, 可以提高其后继结点 (包含后继阻塞点) 输出能力, 从而提高其吞吐率, 至少可以延缓阻塞传染。但是如果关键阻塞结点存在前驱阻塞结点, 它是可以利用阻塞路径集合得到。由于它对首阻关键结点有较大的影响, 因此应与首阻关键结点等同时处理。

如果集合 Y 为空, 表明网络只是局部阻塞。但是从算法可知结点 i 是关于路径集合 Z_i 的关键阻塞点, 因此我们必须求出关于最大阻塞路径数的阻塞结点, 即求取最大的阻塞子网。根据算法 A-4.1, 显然有如下结论。

定理 4.3 根据算法 A-4.1, 如果阻塞结点 $v_i \in N$, 具有最大的关联系数值 q_i , 则相应的 Z_i 所对应的子图为关于阻塞结点的最大阻塞子图。

假设最大阻塞子图的关键阻塞点及其首阻点阻塞状态的解除, 其子图中的其余结点的阻塞条件同时解除。那么再次按算法 A-3.1 和算法 A-4.1 计算图的阻塞路径和阻塞结点, 从而确定关键阻塞点和最大阻塞子图, 以解除其阻塞状态。反复

进行即可将网络恢复到无阻塞或阻塞子图最小的状态。

五、算法的应用及实现

至此, 我们可以通过阻塞路径的求取, 进而求得关键的阻塞结点, 为网络的强化管理和控制提供了一种手段。尽管网络业务及流量的不断变化, 特别是网络流量分布的变化, 使得相应网络的关键阻塞结点 (集) 也随之发生变化, 但是关键阻塞点和阻塞现象在一定时期内具有相对稳定的表现形态。

通常情况下, 阻塞具有时效性, 且多发生在业务的高峰期, 关键阻塞结点的阻塞解除, 比其它结点的阻塞解除, 更有利提高网络的运行效率, 带来更多的收益。通过电信网管系统对实际运行的电信本地网的状态进行统计分析的结果表明, 网络的阻塞状态具有较强的关于时间和业务流量的统计特性, 即绝大多数阻塞状态产生于业务流量的高峰期, 并维持相应的时间。

网络资源和组成结构的改变, 同样会改变关键阻塞结点及分布, 因为阻塞结点同样与网络资源的分布和配置密切相关。对阻塞关键结点的资源进行动态调整和优化, 将大大降低网络阻塞的发生几率。

因此, 可以认为关键阻塞结点是动态的, 而非静态的, 必须根据对网络的不断监测和分析才能及时掌握关键阻塞结点的变化, 这样做不仅提高网络运行效率, 也能为网络的优化提供依据。

所给出算法可以强化网络的管理, 为网络流量的疏导提供了基础。该算法应用及实现的主要步骤如下:

1) 监测网络流量的变化和状态, 实时计算各结点的阻塞度 T' , 并对结点阻塞状态进行概率统计;

2) 设定时间周期, 利用定义 4.1 和算法 A-3.1 进行求取所有的阻塞传染路径;

3) 利用算法 A-4.1 计算出关键阻塞结点;

4) 利用网络结构、结点发生阻塞的时间关系以及概率频度求出网络阻塞的首阻点;

5) 对首阻点和关键结点进行流量控制处理, 解除结点阻塞;

6) 在阻塞结点概率统计的基础上, 可对网络资源及结构的优化进行辅助决策管理。

由此可见, 该算法的应用不仅可以强化网络管理, 特别是在网络结构和资源配置的优化方面, 并为促进网络管理的智能化提供了具体的算法基础; 同时算法的实施也提升了网络对流量的承载能力, 提高了网络的效率。

结论 在前面的描述中我们对阻塞现象进行了分析, 并在求得阻塞传染路径的基础上, 对阻塞结点, 特别是关键阻塞结点以及阻塞路径的阻塞度进行了研究, 提出了求取关键阻塞结点的诊断算法, 并进行了算法分析和论证。由于不同的网型结构和信息转移方式对阻塞传染方式有一定的影响, 依据本文所给出的一些基本思想和概念针对具体的网络, 可以设计出更好的阻塞及传染的测试和诊断算法。

文中所得结论, 对阻塞的解除条件进行分析和研究打下了一定的基础, 但本文所做的工作仅是阻塞传染方面研究的开始, 进一步还可以进行基于阻塞传染的资源再配置技术和流量平衡等方面的研究。

参考文献

- 1 黄开源. 系统诊断的传染故障模型. 重庆大学学报, 1985, (5): 49

基于最小风险的贝叶斯邮件过滤算法^{*}

Mail Filtering Based on the Risk Minimization Bayes

石霞军 林亚平 陈治平

(湖南大学计算机与通信学院 长沙 410082)

Abstract A new algorithm of filtering junk-mail based on the risk minimization Bayes is proposed in this paper, which can reduce the error rate of misclassifying legal mail. The experiment results show that the algorithm has better performance.

Keywords Feature extraction, Risk minimization Bayes, Mail filtering

1. 引言

随着因特网的迅猛增长,电子邮件作为最快捷、最经济的通信方式,也得到飞速发展。但是,许多销售广告、快速致富等垃圾邮件也在网络中传送,这些垃圾邮件不仅占据邮件服务器的大量存储空间,同时也要用户花费大量时间来处理这些垃圾。尽管一些商业化产品允许用户人工建立垃圾邮件的过滤规则,但是系统要求用户有丰富经验和花去许多时间,而且,由于垃圾邮件在不断改变,用户必须经常调整这些规则,这需要花大量时间。因此,研究邮件自动过滤方法具有重要意义。邮件自动过滤方法研究主要有基于规则和基于概率^[1]两种,后者已成为一种主要研究趋势。

文[5]通过动态学习规则将邮件进行分类,这种方法对基于邮件内容进行一般分类时具有一定作用,但在邮件过滤过程中没有考虑到垃圾邮件本身具有与合法邮件不一样的特性,而且文中提出的 RIPPER 算法需要消耗大量的时间。

Sahami^[2]利用贝叶斯(Bayes)算法对邮件进行分类,并结合手工构造的短语以及具有垃圾邮件特性的单词来提高过滤垃圾邮件的精确度。但作者为减少合法邮件因被误判为垃圾邮件,将判为垃圾邮件的概率门槛值提高到了 99.9%。这种过高的门槛值造成大部分垃圾邮件被误判为合法邮件的现象。

Rennie^[4]在基于贝叶斯算法的基础上建立了一个用于邮件过滤的机器学习应用系统 ifile,利用贝叶斯算法对邮件进行分类,这种方法分类速度比较快,不需要用户太多干预,错

误分类的邮件可以通过用户对该邮件的不同类间的转移,并利用贝叶斯算法重新训练,可得到动态调整。但是由于错误分类的现象比较多,用户仍需要检查垃圾邮件中是否存在错误分类的合法邮件而要花费大量时间。

本文在朴素贝叶斯算法的基础上,基于最小风险贝叶斯方法提出一种新的邮件过滤算法,在尽量不使合法邮件错判为垃圾邮件的同时,尽可能地提高分类的精确度。实验表明,这种基于最小风险贝叶斯邮件过滤算法具有较好性能。

2. 朴素贝叶斯(Naive Bayes)算法

朴素贝叶斯算法是在一般贝叶斯算法的基础上通过假定各因素之间不存在任何联系,即完全独立而得到的一种简化贝叶斯算法。这种算法在文本分类中得到非常广泛的应用。根据贝叶斯概率公式,对于给定的向量 $d(\omega_1, \omega_2, \dots, \omega_n)$ 属于第 $C_k(k=1, 2, \dots, m)$ 类的概率为:

$$P(C_k|d) = P(C_k) \times P(d|C_k) / P(d) \quad (1)$$

$$\text{其中: } P(d) = \sum_{k=1}^m P(d|C_k) \times P(C_k)$$

由式(1)可知,要判断一个待识别邮件的类别,可以通过计算 $P(C_k|d)$ 概率来完成,它表示出该文档中出现的单词与向量空间模型中特征项的匹配情况,而决定该文档属于第 C_k 类的概率。我们可通过先验概率 $P(C_k)$ 和条件概率 $P(d|C_k)$ 来得到后验概率 $P(C_k|d)$ 。

假定 ω_j 表示第 j 个特征项,基于文档中单词出现的概率相对独立的假设,有:

^{*} 湖南省自然科学基金资助项目(99JJY20060)。石霞军 硕士研究生,研究方向为计算机网络。林亚平 教授,博士生导师,主要研究方向为网络计算和机器学习。陈治平 博士研究生,研究方向为机器学习。

- 2 Bahk S, Zarki M E. Congestion control based dynamic routing in ATM networks. Comput. Commun., 1994, 17(12): 826~842
- 3 Tao yang, et al. Blocking Infection of Network and It's Diagnosing and Analyzing. 2000 Intl. Conf. on Communication Technology Proceedings, IEEE Press (IEEE catalog number: 00EX420) & Publishing House of Electronic Industry, Beijing, China, 2000. 296~301
- 4 Tao yang. The Analysis of the Congesting Control Between the Networks. In: Proc. ISTIST'96 (1996 Intl. Symposium on New Transmission & Switching Technologies), Kun-ming, China, 1996. 318~321
- 5 Tao Yang, et al. The Analysis of Buffer Module Stream Based on

Cross-switch Structure. In: Proc. ISTN'97 (1997 International Seminar on Teletraffic and Network), 1997 Xi'an, China, 1997. 478~481

- 6 Tao Yang, et al. A Study of Throughput Probability of Statistical Multiplexer with Multi-port Buffer Scheme. 信息就是力量('97 香港北京国际计算机会议论文集), 清华大学出版社, 1997. 153~156
- 7 Saluja K K, Anderson B D O. Easily diagnosable design at system level. Conf. on Comput. in Engin., Melbourne: Australia, 1981. 59~63
- 8 陶洋. 异种通信网互连控制关键技术研究: [重庆大学博士论文]. 1998. 6