网络流量有效监测点的设置模型及求解算法研究

蒋红艳 林亚平 黄生叶
(湖南大学计算机与通信学院 长沙 410082)

摘 要 网络流量监测点问题可以抽象为图的最小弱顶点覆盖问题，而求解最小弱顶点覆盖问题是一个 NP 难题。该文利用图论中关联矩阵的概念，提出了一个近似算法，并分析了算法的复杂性。在此基础上将该算法拓展到顶点加权情况下图的弱顶点覆盖问题。理论分析和仿真实验表明，比较现有的算法，新的算法能够发现更小的弱顶点覆盖集，且具有更好的可扩展性。

关键词 图论，弱顶点覆盖，顶点加权，流守恒，NP 难题，关联矩阵
中图分类号：TP393.06, O157.9 **文献标识码**：A **文章编号**：1009-5896(2006)04-0753-04

Model and Algorithm Research for Seeking Efficient Monitor-Nodes Measuring Network Traffic

Jiang Hong-yan Lin Ya-ping Huang Sheng-ye
(*College of Computer and Communication*, *Hunan University*, *Changsha* 410082, *China*)

Abstract The problem of seeking monitor-nodes for measuring the network traffic is regarded as the problem of finding out the minimum weak vertex cover of a graph which is NP-hard. An approximation algorithm is proposed in this paper based on the concept of incidence matrix in Graph. Also the complexity of the algorithm is analyzed. Furthermore, the algorithm is expanded to seek the minimum weak vertex cover for a graph that has weights on the nodes. The theoretical analysis and the simulation results show that the novel algorithm is more scalable than the traditional algorithms, and can find smaller weak vertex cover.

Key words Graph, Weak vertex cover, Nodes with weights, Flow conservation, NP-hard, Incidence matrix

1 引言

随着Internet应用的增长，网管系统越来越注重于应用服务的管理， 网络流量监测需要更大的数据量和更高的数据采集频率。但大量设备的频繁轮询采集增添了网络的额外负担，造成了路由器性能的下降。因此，为了优化网络性能，设置高效合理的流量监测系统就成了重要的研究课题[1]。近年来，一些学者就这一问题进行了研究，研究中一般将网络流量测试问题转化为求给定无向图中的最小顶点覆盖问题或最小弱顶点覆盖问题。由于这些问题都属于NP难题[1,2]，因此一般采用近似求解算法。文献[1,3]提出了一个高效的网络利用率和延迟监控策略及 3 个计算最小弱顶点覆盖的近似算法。文献[2]利用文献[1]中的监控策略从图论角度，对贪心算法进行理论分析，得到了该算法的比界和时间复杂度。但这些算法都没有考虑顶点加权的情况。本文在文献[1]的基础上利用关联矩阵和线性规划的概念，提出了一个求最小弱顶点覆盖问题近似算法，并将该思路推广到顶点加权情况。文中分析了该算法的时间复杂性；并通过仿真实验比较了文献[1]提出的算法，结果表明，该算法可以找到更小的覆盖，且容易推广到顶点加权的情况，因此具有更好的可扩展性。

2004-11-29 收到, 2005-05-19 改回
湖南省自然科学基金(03JJY3098)资助课题

2 预备知识

在网络中某一节点设置监测器(如 SNMP Agent)，可以得到与这一节点相联的所有链路上的流量。因此，为了得到网络中所有链路的网络流量，一般可以通过在某些交换节点(路由器)配置监测器实现。我们要考虑的问题是：在网络哪些节点上设置监测器，才能使得在可以得到每一条链路流量的条件下，所需流量监测器数目最小。这一问题可以归结为图论中求最小顶点覆盖问题。

典型的网络结构可以描述为一个无向图 $G=(V,E)$，其中 $V=(v_1,v_2,\cdots,v_n)$ 为网络节点集(在 IP 网络中可看作为路由器)， $E=(e_1,e_2,...,e_m)$ 为网络中的链路集合。其中， $n=|V|$，$m=|E|$ 分别表示 G 中节点和链路的数目。用 $e_k=(v_i,v_j)$ 表示 e_k 是连接节点 v_i 和 v_j 的一条链路，用 Degree(v) 表示节点 v 的链路数。参照文献[2]，给出以下定义。

定义 1(最小顶点覆盖问题) 最小顶点覆盖问题，是指给定一个无向图 $G=(V,E)$，求顶点集 V 的一个最小子集 S，使得 $e=(u,v)\in E$，且 $u\in S$ 或 $v\in S$，即 E 中的任一边至少含有此子集中的一个点作为顶点，也就是说 S 中的顶点覆盖了边集 E。$u\in S$ 或 $v\in S$。

从定义 1 易见，若在G的一个覆盖S的每一节点上设置一个监测器，就可以得到G上所有链路上的流量。有效监测问题的目标就是求解给定图G最小顶点覆盖集S。但已证明，最

小顶点覆盖问题是一个NP难题[2]。如果监测器是路由器或交换机等交换设备，那么图G还满足以下两条约束条件。

(1) 对图 G 的顶点集 V 中的任意顶点 v，其度 $\mathrm{Degree}(v)\ge 2$；

(2) 对图 G 的顶点集 V 中的任意顶点 v，满足流守恒方程，即流入＝流出。

尽管以下原因会导致流守恒方程的失真，如：(1) 交换设备是数据的源或汇，而不仅仅是数据转发器；(2) 多播导致输出端口的数据复制；(3) 交换设备本身的数据包延迟或丢弃。但研究表明，流守恒方程的相对误差小于 0.05% [1,4]。

定义 2(弱顶点覆盖) 假定无向图 $G=(V,E)$ 满足对任意 $v\in V$ 有 $\mathrm{Degree}(v)\ge 2$，称 $S\subset V$ 是图 G 的弱顶点覆盖，当且仅当执行以下操作能使 E 中所有边可以被标记。

(1) 标记所有与 S 中顶点相关联的边；

(2) 若某个顶点 v 的 $\mathrm{Degree}(v)-1$ 条邻边已被标记，则标记剩下的那条邻边；

(3) 重复第(2)步直到不能标记新的边为止。

显然，由此弱顶点覆盖 S 就可以得到图 G 中各链路的流量。本文利用关联矩阵的概念建立求解问题的模型，为此先给出以下定义。

定义 3(关联关系) 在无向图 $G=(V,E)$ 中，若 $v\in V$ 是 $e\in E$ 的顶点之一，则称 v 与 e 之间存在着关联关系，记为 $v\mathrm{Re}$。

定义 4(关联矩阵) 图 $G=(V,E)$ 的关联矩阵 $\boldsymbol{A}=(a_{ij})$ 是指如下定义的 $n\times m$ 矩阵:

$$a_{ij}=\begin{cases}1, & v_i\,\mathrm{Re}_j\\ 0, & \text{其他}\end{cases}$$

如图 1 中的 G 为某网络拓扑图:

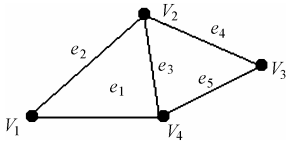

图1 流量监测点选择示意图

Fig.1 Network graph G and efficient monitor-nodes measuring network traffic

对于图 1 中的 G，其关联矩阵 $\boldsymbol{A}$ 为

	e_1	e_2	e_3	e_4	e_5
v_1	1	1	0	0	0
v_2	0	1	1	1	0
v_3	0	0	0	1	1
v_4	1	0	1	0	1

根据关联矩阵的定义，我们知道 V 的一个子集构成图 G 的一个覆盖，当且仅当在它包含的节点所对应的关联矩阵的行中每列至少存在一个 1。

3 有效监测点的设置模型和选点算法

由以上分析，有效监测点设置数目最小的问题可转化为以下最小覆盖问题，令：

$$x_i=\begin{cases}1, & v_i\in S\\ 0, & v_i\notin S\end{cases}$$

则问题变为求： $\min\sum x_i$ 且满足：对每个 $j\in\{1,2,\cdots,m\}$，$\sum_{j=1}^{m}a_{ij}\ge 1$， 其中 a_{ij} 为关联矩阵中对应的元素。

基于这种考虑，下面我们给出一个选择最小覆盖的算法，算法的思路为:

(1) 选取一个包含的链路数目最多的节点，记为 v_i；

(2) 删去关联矩阵中 v_i 对应的行及该行中值为 1 的元素所在的列；然后在剩下的关联矩阵中再依次删除所有行元素之和不超过 1 的其他行及这些行中值为 1 的元素所对应的列，直到不能再删除新的行和列为止；

(3) 重复以上步骤的操作，直到所有的链路都被包含到。

为方便起见，称之为 Select−Measured−Nodes− Algrithm (简称 SMN)，具体算法如下:

输入：网络拓扑图 $G=(V,E)$，其中 $V=(v_1,v_2,\cdots,v_n)$，$E=(e_1,e_2,\cdots,e_m)$；

输出：测量节点集合 S

Begin

(1) 写出图 $G=(V,E)$ 的关联矩阵 $A_1=A=(a_{ij})$，其中 $i\in\{1,2,\cdots,n\}$， $j\in\{1,2,\cdots,m\}$；

(2) 令 $S=\varnothing$，k=1；

(3) while (A_k 非 0)；

Begin

(4) $\forall\ i\in(1,2,\cdots,n)$,计算关联矩阵 A_k 的每行元素之和 $\sum_{j=1}^{m}a_{ij}$，选取和为最大的那一行记为 $\max_i\sum_{j=1}^{m}a_{ij}=a_k$，若 A_k 中有两行i_1，i_2的行和都对应于最大值 a_k，则当 $i_1\le i_2$ 时，取 $k=i_1$，即得到相应于该行的顶点 v_k；

(5) $S=S+\{v_k\}$;

(6) 删去关联矩阵 A_k 中 v_k 对应的行及该行中值为 1 的元素所在的列；然后在剩下的关联矩阵中再依次删除所有行元素之和不超过 1 的其他行及这些行中值为 1 的元素所对应的列，直到不能再删除新的行和列为止，令 A_{k-1} 为最后得到的关联矩阵；

(7) k=k+1；

End // end of while ()

End // end of Algrithm

以图 1 为例，我们对算法 SMN 说明。

e_1 e_2 e_3 e_4 e_5

(1)图 1 中图 G 对应的关联矩阵

$$\boldsymbol{A}_1=\begin{array}{c}v_1\\v_2\\v_3\\v_4\end{array}\begin{pmatrix}1&1&0&0&0\\0&1&1&1&0\\0&0&0&1&1\\1&0&1&0&1\end{pmatrix}_{4\times5}$$

(2)当 i=1 时，$\sum_{j=1}^{5}a_{1j}=2$；当 i=2 时，$\sum_{j=1}^{5}a_{2j}=3$；当 i=3 时，$\sum_{j=1}^{5}a_{3j}=2$；当 i=4 时，$\sum_{j=1}^{5}a_{4j}=3$；可知当 i=2 时，行和值最大，所以选取 v_2；

(3)删去关联矩阵中 v_2 对应的行及该行中值为 1 即 e_2，e_3 和 e_4 所对应的列后，得到剩下的关联矩阵

$$\begin{array}{c}\\v_1\\v_3\\v_4\end{array}\begin{array}{c}\begin{matrix}e_1&e_5\end{matrix}\\\begin{pmatrix}1&0\\0&1\\1&1\end{pmatrix}_{3\times2}\end{array}$$

在这个关联矩阵中再依次删去所有行元素之和不超过 1 的其它行如 v_1，v_3 及这些行中值为 1 的 e_1 和 e_3 所对应的列，得到 $A_2=0$。于是我们得到该算法 SMN 所得的覆盖集为 $S=\{v_2\}$。即只要轮询一个节点即可得到各链路的流量，而按照常规的测量方法需要轮询全部的 4 个节点。

下面分析该算法所具有的性质：

定理 1　若图 $G=(V,E)$ 是简单连通无向图，且满足对任意 $v\in V$ 有 Degree(v)≥2，则 SMN 算法所得到的节点集 S 是图 G 的一个弱顶点覆盖集。

证明　对图 $G=(V,E)$ 的顶点数 $n=|V|$ 用数学归纳法。因图G满足对任意 $v\in V$ 有 Degree(v)≥2，故 $n\geq3$。当 $n=3$ 时，显然定理 1 成立。假设定理 1 对 $3\leq n\leq k$ 时成立，下面证明定理 1 对 $n=k+1$ 时也成立。在图G的顶点度数最大的节点 v_1 上设置一个监测器，则可以得到与节点 v_1 相关联的所有链路上的流量，将这些链路作上标记，并将节点 v_1 纳入节点集S_1中(等价于将这一点及与该点相关联的链路删除)。考虑剩下的网络拓扑图 $G'=G-v_1$。这时根据上述SMN算法中的步骤(6)可知 $3\leq|V(G')|\leq k$ 或 $|V(G')|=0$。若 $|V(G')|=0$，则 $S_1=\{v_1\}$。由弱顶点覆盖集的定义可知 $S_1=\{v_1\}$ 是一个弱顶点覆盖集。若 $3\leq|V(G')|\leq k$，仍由SMN算法步骤(6)可知图 G' 满足对任意 $v\in V(G')$ 有 Degree(v)≥2。由数学归纳法假设，对 G' 用SMN算法得到的节点集 S' 是一个弱顶点覆盖集。又根据S_1的选取和弱顶点覆盖集的定义可知 $S=S'\cup S_1$ 是一个弱顶点覆盖集。　证毕

定理 2　对 SMN 算法中的图 G 和 S，若用 $G'=G-S=\{G_1,G_2,\cdots,G_j\}$ 表示在图 G 中去除 S 及相邻边的集合，$G_I(1\leq I\leq j)$ 表示 G' 中的各连通的子图，N_i 和 M_i 分别表示 G_I 中的节点数和边数，则对任何给定的 G_I，有 $N_i\geq M_i>M_i$。

证明　当 V_i 是 G_I 中的任何顶点时，有

$$\sum_{e_i\in Iv}Bw(e_i)-\sum_{e_j\in O_v}Bw(e_j)=0 \qquad (1)$$

其中 I_v 为节点 v 的所有输入链路，O_v 为节点 v 的所有输出链路，则可以列出含 N_i 个流量守恒方程的方程组。该方程组中的变量为链路流量，变量个数为 M_i，易知对任何一个给定的 G_I，当 $N_i\geq M_i$ 时，方程组式(1)有解。即 G 中各条链路流量可以被测量和推导。反之，若存在一个 G_I，它相应的 N_i，M_i 满足 $N_i<M_i$，则 G 中各条链路流量不可推导。即有 $N_i<M_i$ 条链路无法被标记。这与弱顶点覆盖集定义相矛盾。所以有 $N_i\geq M_i$。　证毕

定理 3　在输入网络拓扑图为 $G=(V,E)$ 时，其中 $\boldsymbol{V}=\{v_1,v_2,\cdots,v_n\}$ 为网络节点集，$\boldsymbol{E}=\{e_1,e_2,\cdots,e_m\}$，设 $t=\max(m,n)$ 时，上述 SMN 算法的时间复杂性为 $O(t^2)$。

证明　假设上述SMN算法所得到的节点集$\boldsymbol{S}$是用链表L表示，图 G 用 $n\times m$ 阶关联矩阵$\boldsymbol{A}$表示，则SMN算法中语句(1) 的时间复杂性为 $O(t^2)$；语句(2)需常数时间；语句(4)计算各节点的度 $d(i)$ 需时间 $O(t)$；语句(5)需常数时间；语句(6)因为每次循环至少删去一个点，所以循环次数最多为 $O(t)$；删去度 $d(i)$ 最大的顶点 v_i 及与 v_i 相关链的边(即删去关联矩阵中行值最大的行及该行中值为1相对应的列)需时间 $m+d(i)\times(n-1)$，即时间复杂性为 $O(t^2)$；在剩下的 $(n-1)\times(m-d(i))$ 的关联矩阵中删去一个度为1的行及该行中值为1相对应的列需时间 $(n-1)+(m-d(i))$，时间复杂性为 $O(t)$。因此，SMN算法的时间复杂性为 $O(t^2)$。　证毕

由定理 3 可知，本文SMN算法的时间复杂性为 $O(t^2)$，它与文献[1,5]中的最大匹配算法的时间复杂性差不多，但比其中Greedy Rank算法[1]的时间复杂性 $O(t^3)$ 要好。

下面我们将算法推广到图 G 顶点加权的情况。假设在节点 v_i 处设置有效监测点的费用为 f_i，显然 f_i 越大，在 v_i 处设置监测点的价值越小，v_i 被选中的可能性就越小。另一方面，在节点 v_i 处设置监测点的投资价值可用经过此节点的流量和其被访问的频率来度量，我们用 b_i 表示。它可通过统计软件来进行测量，b_i 具有一定的稳定性，但也可能因为各种因素影响而呈现动态变化。因此动态调整 b_i 的大小，可以动态调整所建立的数学模型及其最优解。显然 b_i 越大，在 v_i 处设置监测点的价值越大。令 $c_i=f_i/b_i$，可看出 c_i 给出了在 v_i 处设置监测点的价值的一个量的描述，令 $d_i=c_i\Big/\sum_{j=1}^{m}a_{ij}$，$d_i$ 给出了节点 v_i 处的每条链路的价值，它也是下述算法的一个主要参数。

由以上分析，我们可将上述有效监测点设置价值最高模型转化为以下顶点带权的最小覆盖问题：$\min\sum c_ix_i$ 且满足：对每个 $j\in\{1,2,\cdots,m\}$，$\sum_{j=1}^{m}a_{ij}\geq1$，　其中 a_{ij} 为关联矩阵中对应的元素。

我们知道以上问题也是一个 0-1 规划问题，对这一问题目前尚无较好的算法。我们对 SMN 加以拓广，得到针对顶

点带权的最小覆盖问题的广义 SMN 算法，其算法思路为

(1)选取一个对应的 d_i 最小的节点，记为 v_{i1}；

(2)删去关联矩阵 $\boldsymbol{A}_k$ 中 v_{i1} 对应的行及该行中值为 1 的元素所在的列；然后在剩下的关联矩阵中再依次删除所有行元素之和不超过 1 的其他行及这些行中值为 1 的元素所对应的列，直到不能再删除新的行和列为止，令 $\boldsymbol{A}_{k+1}$ 为最后得到的关联矩阵；

(3)对剩下的节点重复(1),(2)步骤直到最后得到的关联矩阵为 0。

这样，所选取的节点集 $\{v_{i1},v_{i2},\cdots,v_{in}\}$ 即为覆盖集。

仍然以图 1 为例对广义SMN算法的思想加以具体说明。不妨设$(f_1,f_2,f_3,f_4)=(20,30,20,54)$, $(b_1,b_2,b_3,b_4)=(2,5,4,6)$，则可以得到$(c_1,c_2,c_3,c_4)=(10,6,5,9)$，$(d_1,d_2,d_3,d_4)=(5,2,2.5,3)$，我们可以用图 2 表示：

d_i	c_i		e_1	e_2	e_3	e_4	e_5
5	10	v_1	1	1	0	0	0
2	6	v_2	0	1	1	1	0
2.5	5	v_3	0	0	0	1	1
3	9	v_4	1	0	1	0	1

图 2

Fig.2

因为d_2最小，选取v_2；删去关联矩阵 $\boldsymbol{A}_1$ 中的 v_2 行及 e_2，e_3 和 e_4 列，得到剩下的矩阵：

d_i	c_i		e_1	e_5
10	10	v_1	1	0
5	5	v_3	0	1
4.5	9	v_4	1	1

再依次删去所有行元素之和不超过 1 的其它行如 v_1，v_3 以及这些行中值为 1 的 e_1 和 e_5 所对应的列，得到 $\boldsymbol{A}_2=0$。这样得到覆盖集为 $S=\{v_2\}$。

4 仿真实验

我们参照文献[1]进行仿真实验，比较了几种选点算法的性能。实验中所用到的网络拓扑结构通过 Waxman 网络模型[4]生成。Waxman 网络模型是网络研究中一种比较流行的拓扑模型，通过设置它的 3 个参数，我们可以得到各种不同的网络拓扑结构：(1) n，即拓扑图中所含节点个数；(2) α 用于控制拓扑图中边密度的参数；(3) β 用于控制拓扑图中节点平均度数的参数。

我们比较了 4 种算法：生成简单顶点覆盖(Simple VC)的最大匹配启发算法) [1,5]，生成弱顶点覆盖(WVC)的最大匹配算法[1,5]，生成弱顶点覆盖的Greedy Rank算法[1]，以及本文的 SMN 算法的性能。我们依序分别用 $N_{\text{match}}^{\text{vc}}$，$N_{\text{match}}^{\text{wvc}}$，$N_{\text{rank}}^{\text{wvc}}$ 及 $N_{\text{smn}}^{\text{wvc}}$ 表示由上述 4 种算法得到的测量点个数；用 Avg. Degree 表示网络拓扑图中节点的平均度数，它随着 β 值的增大而增大；用 $N_{\text{smn}}^{\text{wvc}}/n$ 表示由 SMN 算法所得到的测量点个数与总节点数之比。实验结果如表 1 所示。

限于篇幅，表 1 仅记录了一组实验结果，对于参数的其他设置，我们也得到了类似的结果。由表 1 的结果可以看出，SMN 算法选出点数比前 3 种算法都小。

表 1 4 种选点算法比较

Table1 Comparision of the four algorithms seeking efficient monitor_nodes (n=500, α=0.4, β={0.02,…,0.08})

Avg.Degree	$N_{\text{match}}^{\text{vc}}$	$N_{\text{match}}^{\text{wvc}}$	$N_{\text{rank}}^{\text{wvc}}$	$N_{\text{smn}}^{\text{wvc}}$	$N_{\text{smn}}^{\text{wvc}}/n$
4.4	387	255	165	129	0.26
8.6	441	372	254	221	0.44
12.6	453	408	307	296	0.59
16.9	466	431	334	327	0.65

5 结束语

本文对网络流量有效监测点问题提出了一个近似算法，该算法可以推广到求解顶点加权情况下图的弱顶点覆盖问题。与文献[1,5]中的算法相比，本文的算法具有更好的性能。对于大规模复杂网络的有效点监测优化设置问题，每选取一个优化节点，可以很大程度降低问题的规模，达到优化网络性能的目的。虽然本文考虑的是基于集中控制的网络系统，即通过设置一定范围的一个网络中心，按一定频率主动收集各网管代理的信息而测得网络流量。但通过将大网络划分为若干子网络，而各子网络可由集中控制求得各自的网络流量，再将大网络看成是由各子网络中心作节点组成的网络，就可以再次运用本文的思想求得整个网络的流量。由此将集中控制推广到分散控制。这也是我们下一步工作中要研究的内容。

参 考 文 献

[1] Breitbart Y, Chan Chee-Yong, Garofalakis M, Rastogi R, Silberschatz A. Efficiently Monitoring Bandwidth and Latency in IP Networks. Proceedings of IEEE INFOCOM 2001, Anchorage, Alaska, April 2001, vol.2: 933–942.

[2] 刘湘辉, 殷建平, 唐乐乐, 赵建民. 网络流量的有效测量方法分析. 软件学报, 2003, 14(2)：300–304.

[3] Vazirani V V. Approximation Algorithms. Berlin, Springer-Verlag, 2001: 93–129.

[4] Caceres R, Duffield N G, Feldmann A, *et al*.. Measurement and analysis of IP network usage and behavior. *IEEE Communications Magazine*, 2000, 38(5): 144–151.

[5] Waxman B M. Routing of multipoint connections. *IEEE J. on Selected Areas in Communications*, 1988, 6(9): 1617–1622.

蒋红艳：女，1966 年生，博士生，研究方向为计算机通信网络.

林亚平：男，1955 年生，博士，教授，博士生导师，主要研究方向为计算机网络、机器学习.

黄生叶：男，1966 年生，博士，副教授，主要研究方向为计算机通信网络和分布式计算.