网络延迟主动测量结果的被动测量校准方法

蔡志平, 殷建平, 刘湘辉, 吕绍和, 刘 芳
(国防科技大学计算机学院, 湖南长沙 410073)

摘 要: 网络延迟是提供 QoS 保证、监控和优化网络性能的重要指标. 测量网络延迟主要采用主动测量和被动测量这两种测量方法, 但是这两种方法都存在一定的缺陷. 以主动测量获得的延迟作为用户数据包延迟的一个估计量, 用被动测量设备检测到的用户数据包信息对主动测量数据进行校准, 可以实现一种有效的网络延迟测量方法. 这种方法具有与协议无关、对网络流量影响小、测量方法简单、可测量单个用户流等优点. 我们综合考虑了探测包间用户数据包数量和相邻探测包延迟的变化, 比以往的算法更准确地反映了网络的实际状况, 特别是在网络拥塞甚至存在丢包的情况下, 误差更小, 测量结果更精确.

关键词: 网络监测; 延迟测量; 主动测量; 被动校准
中图分类号: TP393.1 文献标识码: A 文章编号: 0372-2112 (2005) 11-1929-04

Measuring Network Delay Using Active Probe with Passive Calibration

CAI Zhi-ping, YIN Jian-ping, LIU Xiang-hui, LV Shao-he, LIU Fang
(*School of Computer, National University of Defense Technology, Changsha, Hunan 410073, China*)

Abstract: Latency is a key performance parameter and utilization indicator for providing QoS guarantees and monitoring network performance. In general, monitoring schemes to measure latency are, divided into two types, active and passive monitoring. Unfortunately, both types have drawbacks. The method of combining passive and active approaches uses the latency gained by active measurement as an estimator of the latency of user packets, with weight adjusted by the number of user packets obtained by passive monitoring. This effective method has some advantages such as protocol-independent, negligible extra traffic, convenience and being able to estimate individual user performance. This method considering the number of user data packets arriving between probe packets and the latency alteration of neighborhood probe packets, could reflect the actual network status more exactly, especially in the case of network congestion and packet loss.

Key words: network monitor; measuring latency; active measurement; passive calibration

1 引言

网络延迟是现代网络的重要指标, 它对提供 QoS 保证、通信工程、错误和拥塞检测、网络性能调试、网络管理都有非常重要的指标作用[1~3]. 因为带宽测量[4]、丢包率测量[5]、流统计[6]甚至提供 QoS 保证[7,8]都与延迟的测量有关, 所以对延迟测量精度的要求也越来越高.

测量网络延迟主要采用主动测量和被动测量两种方式. 主动测量方式通过发送测量包来获取链路的延迟; 被动测量方式使用接入网络的探针来记录和统计链路上数据包的网络特性. 但是这两种都存在一定的缺陷[9,10].

当使用测量用的数据包来充当实际的用户流时, 增加的测量数据流量会影响网络的性能, 从而影响用户对网络的使用; 另一方面, 通过测量包获得的网络性能指标并不能等同于没有测量数据流影响下的网络实际性能. 被动测量方式可以通过两点监控或者单点监控来完成. 两点监控测量方法要求所有的设备必须时间同步, 而且因为需要查看包头或者数据包内容来判断是否为同一个包, 网络流量越大, 从大量数据包中过滤和鉴别数据包对系统性能的影响也会越大. 在大规模网络中, 两点监控的被动测量方法不具备很好的可扩展性. 单点监控利用了 TCP 的确认机制, 但仅限于检测 TCP 数据流.

Masaki Aida 提出了一种基于测量变换的结合主动和被动测量方式的网络延迟测量方法 CoMPACT[9,10]. 他们通过少量的主动探测包获取测量数据, 作为用户包数据的一个估计量, 再用被动测量得到的用户数据包数量信息, 进行似然率的调整, 最后得到无偏的测量结果. CoMPACT 方法对网络性能影响小, 可扩展性强, 并且可以分别得到单个用户、组织和应用的性能数据.

CoMPACT 方法假设相邻两个探测包之间的时间间隔非常短, 并认为在这个时间间隔内网络性能的变化可以忽略不计. 即使发送探测包的间隔非常小, 在网络拥塞甚至大量丢包的情况下, 被动测量设备检测到的探测包之间的间隔可能会增大, 从而给评估结果带来较大误差. 我们分析了相邻数据包延迟变化之间的关系, 提出了主动测量数据的被动校准方法. 我们的方法不但考虑了用户数据包信息, 还分析了相邻探测包延迟之间的关系, 在相同的测量数据基础上, 可以得到比

收稿日期: 2004-07-22; 修回日期: 2005-07-07
基金项目: 国家自然科学基金(No. 60373023)

CoMPACT 方法更为精确的评估结果,更能反映网络的实际情况.

2 利用用户包数量校准主动测量数据

定义测量对象链路延迟为 X, 测量时间区间为 $[0, T]$, 链路延迟函数定义为 $D(t)$, 指标函数 ϕ 定义为 $\phi(t, a)=\begin{cases}1, & D(t)>a\\ 0, & D(t)\leqslant a\end{cases}$, 那么对于任意 $a>0$, 链路延迟的概率分布函数 $Pr(X>a)=\dfrac{\int_0^T \phi(t, a)\,dt}{T}$.

若在测量期间通过链路检测点的用户数据包有 n 个, 探测包有 m 个, 并且第 i 个数据包或探测包的延迟为 A_i, 指标函数 ϕ 定义为 $\phi(i, a)=\begin{cases}1, & A_i>a\\ 0, & A_i\leqslant a\end{cases}$, 则根据用户数据包得到的链路延迟的概率分布函数 $Pr(X>a)=\dfrac{1}{n}\sum_{i=1}^{n}\phi(i, a)$, 平均延迟 $M_u(X)=\dfrac{1}{n}\sum_{i=1}^{n}A_i$. 根据探测包得到的链路延迟的概率分布函数 $Pr(X>a)=\dfrac{1}{m}\sum_{i=1}^{m}\phi(i, a)$, 平均延迟 $M_m(X)=\dfrac{1}{m}\sum_{i=1}^{m}A_i$.

直接测量用户数据包的延迟是很困难的, 因为不但需要测量设备时间同步, 而且在大流量的数据流中过滤、分析数据包也会影响性能. 因为探测包一般比用户包要小, 数量也远小于用户数据包的数量, 所以用探测包获得的网络性能并不是用户直接感受到的网络性能.

虽然不能直接获得用户数据包的链路延迟, 但是当探测包的发送间隔 Δt 足够小时, 可以假设对于任何一个测量时间点 t, 在 $[t, t+\Delta t]$ 区间内链路延迟的变化非常小, 即有:

$$\forall s, s'\in[t, t+\Delta t)\Rightarrow D(s)\cong D(s') \qquad (1)$$

通过简化的被动测量设备可以获得相邻两个探测包之间到达的用户数据包的数量. 简化后的被动测量设备只需检测探测包的到达, 并具备简单的计数功能. 设被动测量设备检测出第 $i-1$ 个探测包和第 i 个探测包之间, 到达的用户数据包的数量为 ρ_i, 则结合主动和被动测量方式, 测量得到的链路延迟概率分布函数 $Pr(X>a)=\sum_{i=1}^{m}\phi(i, a)\dfrac{\rho_i}{n}$, 平均延迟 $M_c(X)=\dfrac{1}{n}\sum_{i=1}^{m}A_i\rho_i$.

3 利用相邻探测包的测量值变化校准主动测量数据

在网络拥塞的情况下, 以较小的时间间隔发送探测包并不能保证被动测量设备检测到的相邻探测包的间隔也很小, 当探测包到达测量点的时间间隔变大, 甚至探测包丢失时, 都会给测量结果带来误差. 网络拥塞越严重, 探测包丢失越多, 测量结果的误差就越大.

我们可以假设对于任何一个测量时间点 t, 在 $[t, t+\Delta t]$ 区间内链路延迟的变化是连续的. 这样每个探测包的延迟的权重, 不仅与上一个探测包之间到达的用户数据包的数量相关, 而且与上一个探测包的延迟也有关. 由此可以利用探测包间用户包数量和相邻探测包延迟变化来对主动的测量数据进行校准, 我们提出的主动测量的被动校准方法 Pcoam(Passive Calibration of Active Measurement) 的指标函数定义为

$$\phi(i, a)=\begin{cases}1, & A_i>a \text{ 且 } A_{i-1}>a\\ \dfrac{A_i-a}{A_i-A_{i-1}}, & A_i>a \text{ 且 } A_{i-1}\leqslant a\\ \dfrac{A_{i-1}-a}{A_{i-1}-A_i}, & A_i\leqslant a \text{ 且 } A_{i-1}>a\\ 0, & A_i\leqslant a \text{ 且 } A_{i-1}\leqslant a\end{cases}$$

这时, 链路延迟概率分布函数 $Pr(X>a)=\sum_{i=1}^{m}\phi(i, a)\dfrac{\rho_i}{n}$, 平均延迟 $M_p(X)=\dfrac{1}{n}\sum_{i=1}^{m}\dfrac{(A_i+A_{i-1})}{2}\rho_i$, 其中 $A_0=0$.

4 数据包延迟变化分析

包在网络中的延迟可以分为传播延迟、传输延迟、介质访问延迟和排队延迟四个部分. 传播延迟是电信号在链路上传播所需要的时间, 通常为每公里 5 微秒; 传输延迟可由数据包大小与链路带宽的比值得到. 介质访问延迟与网络媒体的特性和网络负载有关, 轻负载时介质访问延迟可以忽略不计, 对于点到点的全双工链路而言, 介质访问延迟为零. 排队延迟是指数据包在路由器中排队所造成的延迟, 它与路由器本身的特性、链路拥塞状况等相关. 因为数据包在路由器中的处理延迟一般为几微秒到几十微秒, 相对于端到端延迟而言可以忽略不计, 所以一般不考虑处理延迟.

传播延迟和传输延迟合称固有延迟, 它反映了端到端延迟的不变部分. 介质访问延迟和排队延迟合称动态延迟, 它反映了端到端延迟的可变部分. 对于要测量的用户流和插入用户流的探测包而言, 因为通过的网络链路相同, 所以介质访问延迟和传播延迟都是一样的. 传输延迟由数据包的大小和链路带宽所决定, 数据包越大, 传输延迟越大, 因为主干带宽一般比较大, 所以传输延迟的差异主要体现在低带宽链路上.

下面分析相邻数据包排队延迟的变化. 先考虑单跳的情况, 根据路由器是否正在处理数据包, 把整个测量时间段分成“忙”期和“闲”期两种状态[11,12], 显然在“闲”期没有数据包在队列中等待, 而在“忙”期队列中等待的数据包的数量大于零, 并且在整个测量时间段中“忙”期和“闲”期是间隔出现的. 设在缓冲区足够大的先来先服务队列中, 第 i 个数据包达到队列的时间为 τ_i, 在队列中的等待时间为 $w_i\geqslant 0$, 接受服务时间为 $x_i>0$, 固定传播延迟 $D>0$, 路由器处理完该包离开的时间为 τ_i^*, 则这个数据包在该跳的延迟 $d_i\equiv\tau_i^*-\tau_i=w_i+x_i+D$, 两个相邻数据包的发送时间之差为 $t_i\equiv\tau_i-\tau_{i-1}$, 到达时间之差为 $t_i^*\equiv\tau_i^*-\tau_{i-1}^*$, 延迟之差 $\delta_{i, i-1}\equiv d_i-d_{i-1}=t_i^*-t_i=(x_i-x_{i-1})+(w_i-w_{i-1})$.

不妨假设第 i 个包为探测包, 下一个探测包为第 $i+k+1$ 个, 即相邻两个探测包之间有 k 个数据包. 当相邻探测包之间间隔比较小时, 可以假设 k 个数据包到达队列的时间是间隔均匀的, 每个数据包接受服务的平均时间为 $\bar{x}=$

$\frac{\tau^*_{i+k+1}-\tau^*_i}{k+1}$. 分析第 $i+h$ 个包, $1\leqslant h\leqslant k$, 它到达队列的时间为 $\tau_{i+h}=\tau_i+h\frac{\tau_{i+k+1}-\tau_i}{k+1}$. 定义函数 $[x]^+=\max(0,x)$, 假定第 i 个包达到队列时, 在队列中等待或正在接受服务的数据包有 p 个, 则第 $i+h$ 个数据包到达队列时在队列中等待或者正在接受服务的数据包有 $\left[p-\frac{\tau_{i+h}-\tau_i}{\bar{x}}+h\right]^+$ 个, 第 $i+h$ 个包的等待时间 w_{i+h} 由式(2)获得.

$$w_{i+h}=\left[p-\frac{\tau_{i+h}-\tau_i}{\bar{x}}+h\right]^+\bar{x} \tag{2}$$

在"忙"期每个数据包的等待时间显然都大于 0, 所以由式(3)得到:

$$\begin{aligned}w_{i+h}&=\left[p-\frac{\tau_{i+h}-\tau_i}{\bar{x}}+h\right]^+\bar{x}=p\bar{x}-(\tau_{i+h}-\tau_i)+h\bar{x}\\&=w_i+h\frac{(\tau^*_{i+k+1}-\tau^*_i)-(\tau_{i+k+1}-\tau_i)}{k+1}\\&=w_i+h\frac{\delta_{i+k+1,i}}{k+1}\end{aligned} \tag{3}$$

而第 $i+h$ 个包在该跳的延迟时间为

$$d_{i+h}=w_{i+h}+x_{i+h}+D=d_i+h\frac{\delta_{i+k+1,i}}{k+1} \tag{4}$$

从式(3)和(4)可以知道在区间 $[\tau_i,\tau_{i+k+1}]$ 中的数据包延迟与该区间两端两个探测包的延迟之差存在线性关系. 因为相邻探测包的间隔 $[\tau_i,\tau_{i+k+1}]$ 较小, 可以假定需要测量的用户流数据包的到达均匀分布在这两个包之间, 所以即使在背景流量存在的情况下, 相邻探测包之间的用户数据包延迟也与区间两端探测包的延迟之差保持线性变化关系.

在多跳情况下, 虽然在各跳中背景流量不同, 但是用户流数据包与探测包的顺序在各跳中保持不变, 其延迟也与相邻探测包的延迟变化保持线性关系. 因为在路由器"闲"期, 式(2)和(3)并不成立, 因此可以设定一个阈值, 当相邻探测包的延迟均大于这个阈值时, 可以假定探测包经过各跳时, 每个路由器都处于"忙"期, 这时使用 Pcoam 方法对这部分数据进行校准.

5　仿真结果

我们采用 ns2 网络模拟器[13]来模拟和评估本文提出的方法, 并以瓶颈路由器的排队延迟为测量对象. 由端到端的延迟减去传播延迟、传输延迟和介质访问延迟, 即可得排队延迟.

Willinger 等人的分析表明[14], 多个具有重尾分布特性的 ON/OFF 数据源相叠加, 可以构造出具有自相似特征的网络环境. 图 1 显示了我们仿真的网络结构. 20 个 ON/OFF 数据源通过 1.5M 的链路与瓶颈路由器相连, 两个路由器之间通过 10M 的链路连接, 所有链路都为 FIFO 队列.

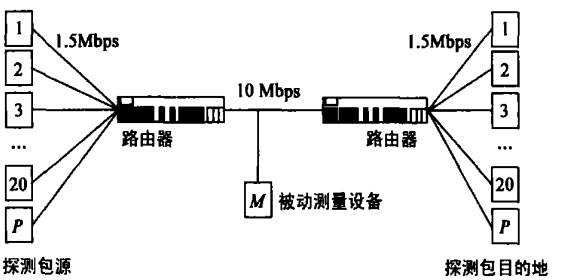

图 1　仿真单跳网络拓扑结构

20 个数据源分成 4 组, 每组 5 个数据发送节点, 分别向对应的数据接收节点发送数据包. 这些发送的数据分别采用 TCP 和 UDP 通信协议, 其 ON/OFF 时间间隔分别服从 Exponential 和 Pareto 分布, 如表 1 所示.

表 1　单跳拓扑仿真数据源参数配置

Node	Protocol	Packet length	Mean ON period	Mean OFF period	ON/OFF length distribution	Shape parameter	Rate at ON period
# 1~ # 5	TCP	1.5KB	10s	5s	Exponential	—	1Mbps
# 6~ # 10	UDP	1.5KB	5s	10s	Exponential	—	1Mbps
# 11~ # 15	TCP	1.5KB	10s	10s	Pareto	1.5	1.5Mbps
# 16~ # 20	UDP	1.5KB	5s	10s	Pareto	1.5	1.5Mbps

探测源节点每隔 2 秒向探测数据接收节点发送一个探测包, 大小为 64 个字节. 在两个路由器之间设立一个被动测量设备, 计算相邻探测包之间用户数据包的数量. 整个测试时间为 1800 秒, 共产生 1931369 个数据包, 其中包括 900 个探测包. 因为探测包的大小固定为 64 个字节, 所以探测包占网络总流量的 0.0047%, 占主干带宽的 0.00256%. 图 2 给出了通过探测包获得的排队延迟, 从图中可以看出通过探测包获得的排队延迟的抖动比较大.

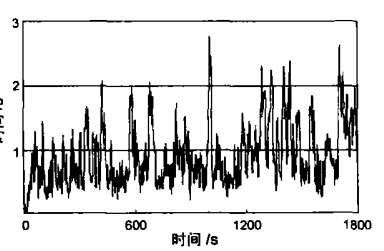

图 2　通过探测包获得的排队延迟

我们从 20 对节点中, 选取了节点 6 来进行分析. 这对节点之间采用 UDP 通信协议, ON/OFF 分布服从 Exponential 分布. 图 3 和图 4 中从上到下的 4 条曲线依次表示根据 CoMPACT 方法、数据包方法、Pcoam 方法和探测包方法获得的排队延迟概率分布, 图 4 在 y 轴上采用了对数刻度, 可以更清楚地分析小概率事件的分布. 从图中可以看出, 探测包的延迟并不能完全代表用户数据包的延迟, 而我们提出的 Pcoam 方法, 无论是整体概率分布, 还是小概率事件分布, 都比 CoMPACT 方法更接近用户包的实际延迟. 因此相对于 CoMPACT 方法, 使用主动测量数据的被动测量校准方法, 更能反应网络的实际状况.

表 2 中列出了用 4 种方法分别计算得到的 4 类节点的平均延迟. 从表中可以看出, 采用主动探测包的主动测量方法只能得到网络性能的大致状况, 不能反映出单个用户数据流的特性. 采用主动测量数据的被动测量校准方法, 比 CoMPACT 方法更准确地反映了单个用户数据流的状况.

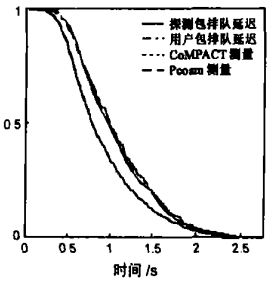

图 3 排队延迟的概率分布 $Pr(X>a)$

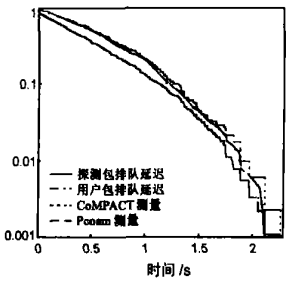

图 4 排队延迟的概率分布 $Pr(X>a)$(对数刻度)

表 2 各种测量方法得到的平均延迟

节点	用户数据包平均延迟(毫秒)	探测包平均延迟(毫秒)	CoMPACT 方法平均延迟(毫秒)	Pcoam 方法平均延迟(毫秒)	用户包数量
# 1	724.746	910.094	741.219	735.108	46298
# 6	1048.390	910.094	1072.290	1038.170	44048
# 11	744.670	910.094	761.245	755.633	43458
# 16	949.759	910.094	959.408	947.159	176703

因为 ns2 模拟网络流量的产生具有一定的随机性,我们选取了 50 组实验结果计算平均值和方差.计算结果表明通过 Pcoam 方法得到的平均延迟要比 CoMPACT 方法更接近于用户数据包平均延迟,而两者的方差则相差不大.我们设定用户数据量的多种流量模式,在仿真单跳和多跳网络拓扑中均进行了多次的测量和计算.实验结果表明主动测量数据的被动校准方法能够更准确地反映网络性能.

6 结束语

主动测量数据的被动校准方法,可以克服单独使用主动或被动测量方式的缺陷,并且具有测量方法简单、对网络影响小、与协议无关、可扩展性强、可测量单个用户流等优点.主动测量数据的被动校准方法,综合考虑了探测包间用户数据包信息和相邻探测包之间的关系,能得到比 CoMPACT 方法更精确的测量结果,使测量结果更能反映网络的实际状况.特别是在网络拥塞甚至存在丢包的情况下,主动测量数据的被动校准方法比 CoMPACT 方法更合理,误差更小,改进效果也更明显.这种方法在 IP 网络或者虚拟专网的链路测量中都可以得到很好的应用.

我们采用模拟实验的途径验证了这种方法可以有效地测量链路延迟.这种方法也可以扩展到测量网络带宽、丢包率和抖动.如何利用尽量少的测量信息得到足够准确的测量结果是我们继续研究的内容.

致谢 作者对 Masaki Aida 博士的鼓励和帮助表示衷心感谢.

参考文献:

[1] G Almes, S Kalidindi, M Zekauskas. A one-way delay metric for IPPM [S]. RFC2679, 1999.

[2] Breibart Y, Chan C Y, Carofalakis M, Rastogi R, Silberschatz A. Efficiently monitoring bandwidth and latency in IP network[A]. INFOCOM 2001[C]. Anchorage, USA, 2001. 939-942.

[3] 张宏莉,方滨兴,胡铭曾,姜誉,张树峰. Internet 测量与分析综述[J]. 软件学报, 2003, 14(1): 110-116.
Zhang Hong-li, Fang Bin-xing, Hu Ming-zeng, Jiang Yu, Zhang Shu-feng. A survey on internet measurement and analysis[J]. Journal of Software, 2003, 14(1): 110-116. (in Chinese)

[4] Kevin Lai, Mary Baker. Measuring link bandwidths using a deterministic model of packet delay[A]. ACM SIGCOMM 2000[C]. Stockholm, Sweden, 2000. 283-294.

[5] K Ishibashi, M Aida, S Kuribayashi. Estimating packet loss rate by using delay information and combined with change-of-measure framework [A]. IEEE GLOBECOM 2003[C]. San Francisco, USA, 2003. 3878-3882.

[6] N G Duffield, C Lund, M Thorup. Properties and prediction of flow statistics from sampled packet streams[A]. ACM SIGCOMM Internet Measurement Workshop 2002[C]. Marseille, France, 2002. 159-171.

[7] Kartik Gopalan, Tzi-cker Chiueh. Probabilistic delay guarantees using delay distribution measurement[A]. 12th ACM International conference on Multimedia[C]. New York, USA, 2004. 900-907.

[8] Kartik Goplalan, Tzi-cker Chiueh, Yow-Jian Lin. Delay budget partitioning to maximize network resource usage efficiency[A]. IEEE INFOCOM 2004[C]. Hong Kong, China, 2004. 562-571.

[9] M Aida, N Miyoshi, K Ishibashi. A scalable and lightweight QoS monitoring technique combining passive and active approaches[A]. IEEE INFOCOM 2003[C]. San Francisco, USA, 2003. 125-133.

[10] K Ishibashi, T Kanazawa, M Aida. Active/Passive combination-type performance measurement method using change-of-measure framework [J]. Computer Communications, 2004, E87-B(1): 132-141.

[11] Attila Pasztor, Darryl Veitch. On the scope of end-to-end probing methods[J]. IEEE Communications Letters, 2002, 6(11): 509-511.

[12] Susmit H Patel. Performance Inference Engine (PIE): Deducing more performance using less data[A]. ACM PAM 2000[C]. Hamilton, New Zealand, 2000. 76-84.

[13] UCB/LBNL/VINT network simulator ns (version 2) [OL]. http://www.isi.edu/nsnam/ns.

[14] Willinger W, Paxson V, Taqqu M S. Self-similar and heavy tails: Structural modeling of network traffic[A]. In: Adler R J, Feldman R E, Taqqu M S. eds. A Practical Guide To Heavy Tail: Statistical Techniques and Applications[C]. Boston: Birhauser, 1998. 27-53.

[15] Zhiping Cai, Jianping Yin, Xianghui Liu, Fang Liu, Shaohe Lv. Efficiently monitoring link bandwidth in IP networks[A]. IEEE GLOBECOM 2005[C]. St. Louis, USA, 2005.

作者简介:

蔡志平 男,1975 年 5 月生于湖南省益阳市,2002 年毕业于国防科技大学计算机学院,获计算机科学硕士学位,现为国防科技大学计算机学院博士生,研究兴趣包括网络测量、近似算法.

E-mail: caizhiping_nudt@163.com.

殷建平 男,1963 年 10 月生于湖南省益阳市,1990 年毕业于国防科技大学计算机系,获计算机科学博士学位,现任国防科技大学计算机学院计算机应用技术教研室主任,教授,博士生导师,主要研究方向为算法设计与分析、信息安全、网络测量与模式识别.