Avaliação de Modelos de Predição de Tráfego Utilizando um Analisador de Protocolos *in-line*

Rivalino Matias Jr, Fernando Schutz

Curso de Ciência da Computação - CTTMar – Campus de São José
Universidade do Vale do Itajaí (UNIVALI) - 88.122-000 - São José, SC - Brasil

rivalino@univali.br, fernandoschutz@gmail.com

Abstract. *This paper presents a comparative study of forecast models applied to analyze the traffic from a real wide are network. The instrumentation used for traffic capture was based on a free software solution composed of Linux bridge and ntop, whose flexibility allowed to use the selected sampling strategy. The results shown that the evaluated models, used with success in other knowledge areas, had excellent accuracy (over 99% on average) with highlight for the Holt-winters, linear regression and single exponential smoothing models.*

Resumo. *Este artigo apresenta um estudo comparativo de modelos de predição de séries temporais, os quais foram aplicados na análise do tráfego real de uma rede de longa distância. O instrumental usado para a captura de tráfego foi uma solução de software livre baseada em Linux bridge e ntop, cuja flexibilidade suportou a estratégia de amostragem definida para o trabalho. Os resultados da análise comparativa mostraram que os modelos avaliados, usados com sucesso em outras áreas do conhecimento, em média propiciaram uma excelente acuracidade (> 99%), com destaque para os modelos de Holt-winters, regressão linear e alisamento exponencial simples.*

1. Introdução

Nos últimos anos a engenharia de tráfego (ET), como definida em RFC3272 (2002), tem sido cada vez mais empregada no planejamento e provisionamento de redes de computadores, principalmente em redes de longa distância. Isto se justifica tanto pelo aumento na complexidade da infra-estrutura destas redes, quanto pelos exigentes requisitos das atuais aplicações de software. Neste contexto, a utilização de métodos estatísticos quantitativos para a caracterização e modelagem do tráfego de redes tem ganhado importância crescente nos processos de ET [Brownlee e Claffy 2002].

Os avanços nesta área são observados ao consultar a literatura (ex. [Brownlee e Claffy 2002], [Krishnamurthy *et al*. 2003] e [Papadopouli *et al*. 2006]), onde é crescente a aplicação de novos métodos para a análise e predição de tráfego. Contudo, em termos práticos, a predição de tráfego de redes tem sido utilizada timidamente em ambientes de produção, apesar da disponibilidade de ferramentas para este propósito, principalmente licenciadas como software livre.

Visando contribuir com o corpo de conhecimento nesta área, este trabalho apresenta um estudo de caso envolvendo a mensuração e análise de predição de tráfego de um ambiente real. No tocante à modelagem, o estudo comparou uma coleção de técnicas de predição de séries temporais, objetivando avaliar suas adequações aos dados

de tráfego de uma rede corporativa de longa distância (WAN). A escolha das técnicas avaliadas priorizou aqueles modelos que pudessem ser implementados com relativa facilidade, a fim de testar seu desempenho frente a um cenário real, objetivando identificar potenciais candidatos para utilização no dia a dia do planejamento e gerenciamento de redes. A obtenção dos dados para a composição das séries usadas no estudo foi conseguida usando-se uma solução de análise de tráfego baseada na integração da funcionalidade de Ethernet *bridging*, disponível no *Kernel* Linux, com o analisador de protocolos ntop, ambas tecnologias licenciadas e distribuídas como software livre.

O restante do artigo está organizado como descrito a seguir. A seção 2 apresenta a metodologia utilizada para o desenvolvimento do trabalho. Na seção 3 o ambiente onde o estudo foi realizado é descrito, bem como as tecnologias e o software livre utilizados. A seção 4 faz uma análise dos principais resultados obtidos. Finalmente, na seção 5 são apresentadas as conclusões do estudo, juntamente com algumas considerações sobre projetos de continuidade desta pesquisa.

2. Metodologia

A primeira etapa do trabalho foi obter uma amostra de tráfego para construir a série temporal utilizada na avaliação dos modelos de predição. A estratégia de amostragem considerou um período de coleta de oito semanas, permitindo observar o tráfego típico da rede em dias normais, os comportamentos cíclicos (ex. início e fim de mês) e também os eventos sazonais (ex. datas especiais) em que a rede suporta uma operação diferenciada da empresa em comparação aos demais períodos. A coleta ocorreu de segunda à sexta entre 07:00 e 23:00, visto que nos demais dias e horários não existiu atividade significativa na rede. Estas definições tiveram como base uma amostra piloto obtida e analisada previamente.

Posteriormente à coleta foram removidos os dados discrepantes (*outliers*) da amostra, cuja origem foram problemas de indisponibilidade de sistemas servidores e enlaces de comunicação, bem como a ocorrência de feriados durante o período de coleta. A série final contou com 31 valores representativos do tráfego total (entrada + saída) diário do enlace WAN monitorado.

Os modelos avaliados neste trabalho foram Sen's *slope*, *naive* (*randon walk*) e *naive* ajustado, regressão linear simples, média móvel (ordens 2 e 3), alisamento exponencial simples ($0{,}1 \leq \alpha \leq 0{,}9$) e *Holt-winters*. Uma descrição detalhada destes modelos foge ao objetivo do trabalho e pode ser encontrada em Hanke *et al*. (2001) e Brockwell e Davis (1996). A avaliação dos modelos teve como principal critério a acuracidade das suas predições em relação aos dados observados, sendo utilizada a estatística MAPE (erro percentual absoluto médio) [Hanke *et al*. 2001] para a construção do ranking de classificação. Também, a partir da amostra de tráfego foi possível gerar inúmeras informações que caracterizaram o uso e o comportamento da rede no período observado. Algumas destas informações serão apresentadas na seção 4.

3. Ambiente e Tecnologias Utilizadas

Os dados utilizados neste estudo foram obtidos a partir de amostras de tráfego coletadas de uma rede de longa distância. Esta WAN corporativa possui abrangência estadual e interconecta um escritório central (Matriz) com suas 11 (onze) filiais. A principal função

desta rede é o transporte dos dados do sistema de gestão da empresa (ERP) e arquivos de imagens (fotos digitais) enviados das filiais para a Matriz. Toda a rede é baseada em enlaces *Frame Relay*, sendo que a capacidade do enlace principal da Matriz é de 512 kbps (CIR 256 kbps) e os demais enlaces de 64 kbps (CIR 3 kbps) por filial.

O analisador de protocolos (AP) escolhido foi o ntop [Deri e Suin 2000], principalmente pela sua estabilidade e por oferecer as informações necessárias para este trabalho. Além disso, apesar do estudo ter sido realizado com uma WAN convencional, em futuras pesquisas envolvendo coleta de tráfego de redes de alta velocidade o ntop também poderá ser utilizado, pois já conta com suporte ao PF_RING. Resumidamente, PF_RING é um novo tipo de *network socket* que amplia consideravelmente a velocidade de captura de pacotes em nível de *Kernel* [Deri 2004].

A estratégia adotada para a amostragem do tráfego definiu como local de coleta a entrada/saída do roteador da Matriz. Nesta localização é possível visualizar todo o tráfego que chega e sai tanto da Matriz como de cada filial, visto que não existe comunicação ponto a ponto entre filiais. Na configuração original, o roteador estava diretamente ligado ao *firewall* que por sua vez conectava-se ao *switch* principal da rede interna da Matriz. A inserção do AP entre o roteador e o *firewall* foi adotada, pois as amostras deveriam ser obtidas livres de filtros realizados pelo *firewall*. Alternativas tradicionais como a utilização de um *switch* ou *hub* para a interconexão do roteador com o *firewall*, a fim de derivar o tráfego (ex. *port mirroring*) para o AP, foram consideradas, porém descartadas devido à falta de um *switch* para este propósito.

Portanto, para que a coleta ocorresse de forma transparente optou-se por utilizar uma solução de AP *in-line*. O termo *in-line* usado neste trabalho significa que o AP está diretamente conectado ao nível físico (L1), de forma passiva e sem a mediação de um dispositivo de interconexão, tal como na abordagem utilizada em Fraleigh *et al*. (2001). A solução adotada foi composta pelo ntop executando integrado a uma *bridge Ethernet* localizada entre o roteador e o *firewall*, o que permitiu a amostragem transparente dos dados sem os inconvenientes e custos adicionais da inserção de um *hub*/*switch*. A Figura 1 apresenta esta configuração.

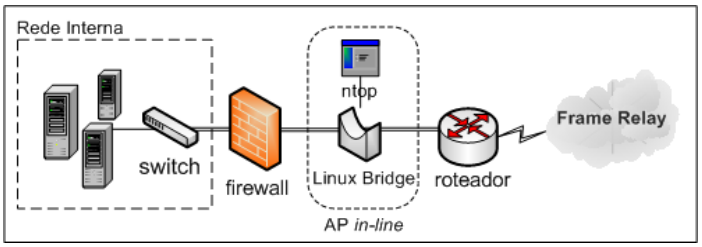

Figura 1. Posicionamento do AP no ambiente de produção

Esta solução foi possível utilizando a funcionalidade de *Ethernet bridging* implementada no *Kernel* Linux e integrada ao ntop. A Figura 2 apresenta os comandos utilizados para a criação da *bridge* Linux e as configurações de carga do ntop.

```
# brctl addbr AP-bridge
# brctl addif AP-bridge eth0
# brctl addif AP-bridge eth1
# ifconfig eth0 up; ifconfig eth1 up; ifconfig AP-bridge up
# ntop -i AP-bridge -d --user ntop -w 192.168.1.244
```

Figura 2: Configurações usadas para a criação da *bridge* e carga do ntop

O projeto Linux *bridge* (bridge.sourceforge.net) adiciona ao *Kernel* Linux a funcionalidade de Ethernet *bridging*, incluindo o protocolo STP (*spanning tree protocol* - IEEE 802.1d), a qual está disponível a partir da versão 2.2 do *Kernel*. O bom desempenho desta funcionalidade pode ser verificado em Yu (2004), o qual apresenta um estudo experimental detalhado comparando o desempenho de uma *bridge* Linux, executando em um PC convencional[1], com um *switch* Cisco Catalyst 2950. O estudo mostrou que para o consumo de CPU da *bridge* Linux inferior a 56%, esta apresentou desempenho comparável ao do *switch* de mercado, correspondendo a um *throughput* de 40.000 fps (*frames per second*). Neste trabalho a configuração adotada para o AP foi: CPU K6-II 550Mhz, RAM 512MB, 2 NICs 100 BaseTX, Fedora Core 4 e ntop versão 3.0.0.1. Esta configuração se mostrou bastante adequada ao tráfego corrente, haja vista que o ntop não reportou nenhuma perda de pacotes durante todo o período de coleta.

5. Análise dos Resultados

Com base no tráfego total diário, do enlace WAN da Matriz, construiu-se a série usada para avaliar os modelos de predição. Vale ressaltar que a avaliação dos modelos ocorreu utilizando uma série "logaritmizada", o que significa que os dados da série original foram transformados ($\log_{10}$) antes da análise. Esta abordagem permitiu um melhor ajuste dos modelos do que com os dados originais. Outras transformações (ln(y), 1/y, e^y) também foram testadas, porém não demonstrando melhores resultados. Dentre os dezesseis modelos considerados neste trabalho (ver seção 2), os cinco primeiros classificados estão listados na Tabela 1.

Tabela 1. Ranking dos modelos aplicados aos dados transformados

Modelos	MAPE	Acuracidade
RLS	0,655102	99,349
H-W	0,720461	99,279
AE(α=0,1)	0,747630	99,252
AE(α=0,5)	0,748280	99,251
AE(α=0,6)	0,752172	99,247

Tabela 2. Ranking dos modelos aplicados aos dados originais

Modelos	MAPE	Acuracidade
H-W	0,720461	99,280
AE(α=0,1)	14,844196	85,156
AE(α=0,4)	15,050288	84,950
AE(α=0,6)	15,068039	84,932
AE(α=0,5)	15,106612	84,893

Os parâmetros do modelo de regressão linear simples (RLS) para a série transformada, obtidos pelo método dos mínimos quadrados, estão apresentados na Equação 1.

$$y = 0{,}0016x + 8{,}5077 \qquad (1)$$

Para efeitos de comparação com a série original, a Tabela 2 apresenta os cinco melhores ajustes para os dados não transformados. Neste caso, o modelo RLS ficou em último lugar (19,355% de acuracidade), motivo pelo qual não apareceu na listagem da Tabela 2. Já os modelos *Holt-winters* (H-W) e alisamento exponencial (AE) apresentaram bons resultados em ambas as séries, com destaque para o desempenho do H-W. Ressalta-se que não foi realizada uma análise exaustiva para a calibração das constantes de amortecimento (α, β, γ) do modelo H-W, tendo estas sido definidas com base em experimentos anteriores.

[1] CPU Duron 1.3 GHz, RAM 512 MB, NIC 100BaseTX

A transformação dos valores das predições para a série original ocorre realizando-se o *anti-log* do valor predito (y). Apesar do modelo RLS ter apresentado a melhor acuracidade para a série considerada na avaliação, a pequena diferença em relação ao segundo colocado é um incentivo para o uso do H-W, visto seu bom ajuste também aos dados originais. Aliado a isso, este modelo tem sido cada vez mais usado (ex. [Brutlag 2000], [Barford *et al*. 2002] e [Krishnamurthy *et al*. 2003]) em pesquisas voltadas para a área de análise e predição de tráfego de redes de computadores. A Figura 3 apresenta uma visão gráfica do ajuste dos modelos RLS e H-W. Os valores do eixo y correspondem ao tráfego total (em megabytes e *logaritmizados*) para cada dia (eixo x) da série analisada.

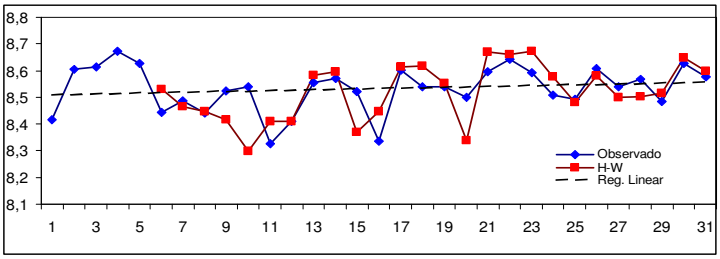

Figura 3. Ajuste dos modelos H-W e regressão linear

Além da análise comparativa dos modelos de predição, a variedade de informações providas pelo ntop permitiu compreender diversos aspectos da utilização da rede tanto da Matriz quanto das filiais. Por exemplo, verificou-se que 97% do tráfego total da Matriz distribui-se entre as aplicações de ERP (39%), transferência de imagem (42%) e acesso web (http/https) (16%). Esta análise foi possível com a customização do ntop para discriminar o tráfego do protocolo proprietário usado pelo ERP. Também, identificou-se que os horários de maior tráfego no período analisado foram de 11:00 até 13:00 e 16:00 até 18:00. Outra análise realizada foi com relação ao sentido dos fluxos de tráfego, onde o maior percentual (61%) foi de tráfego destinado à Matriz. Apesar dos servidores (SQL) do ERP estarem hospedados na Matriz, o tráfego de respostas para consultas e demais dados requisitados pelas filiais foi inferior (39%) ao tráfego de imagens provenientes das filiais para o servidor de arquivos localizado na Matriz.

6. Conclusões e Futuros Trabalhos

A utilização de métodos quantitativos aplicados à caracterização e previsão de tráfego de redes tem se mostrado de grande importância no contexto da engenharia de tráfego. Neste estudo, verificou-se que técnicas tradicionais como RLS e H-W, apesar de simples, apresentaram resultados satisfatórios para as necessidades de um ambiente real como o que foi analisado. Este tipo de constatação serve de estímulo para que estas técnicas possam ser, cada vez mais, aplicadas em situações práticas contribuindo para o aumento na eficácia dos processos de engenharia de tráfego e gerenciamento de redes.

Apesar da larga utilização do modelo H-W em diversas áreas do conhecimento, atualmente suas aplicações em ambientes de produção para a predição de tráfego de redes ainda são consideradas tímidas [Barford et al. 2002], mesmo já tendo sua implementação sido disponibilizada em software livre como é o caso da funcionalidade HWPREDICT do software RRDtool (oss.oetiker.ch/rrdtool). Este software, apesar de oferecer tal recurso, tem sido pouco explorado nesta área, sendo muito usado apenas como repositório de dados. A atual integração do RRDtool com outras plataformas na

área de gerenciamento de redes, como o Nagios (nagios.org) e Cacti (cacti.net), oferece inúmeras oportunidades de aplicações e desenvolvimentos nesta área.

Em termos de futuros trabalhos, dois novos estudos neste campo se encontram em planejamento. O primeiro é a extensão do *plug-in* RRD do ntop, o qual atualmente oferece apenas as funcionalidades de armazenamento e consulta de dados, a fim de torná-lo apto a suportar a técnica de H-W aplicada aos dados capturados pelo ntop. Um segundo projeto é a implementação de outras técnicas de predição no RRDtool que atualmente suporta apenas o H-W. Este projeto inicialmente implementará o método RLS e uma nova funcionalidade para o H-W já implementado, a qual diz respeito à calibração automática das suas constantes de amortecimento. Para isso, pretende-se utilizar como *baseline* uma amostra de uma série armazenada, em conjunto com a estatística MAPE servindo de índice de acuracidade, adotando uma metodologia similar à utilizada neste artigo. A reconfiguração automática dos valores das constantes será realizada com base na funcionalidade rrdtune(1) que é própria do RRDtool.

Referências

Barford, P., Kline, J., Plonka, D. e Ron, A. (2002) "A Signal Analysis of Network Traffic Anomalies", In Proceedings of ACM SIGCOMM Internet Measurement Workshop, França.

Brockwell, P. e Davis, R. (1996) "Introduction to Time Series and Forecasting", Springer.

Brownlee, N. e Claffy, K. (2002) "Understanding Internet traffic streams: Dragonflies and tortoises", IEEE Communications Magazine, 40(10):110--117.

Brutlag, J. D. (2000) "Aberrant Behavior Detection in Time Series for Network Monitoring", In Proceedings of the 14th USENIX Conference on System Administration, New Orleans, USA.

Deri, L. (2004) "Improving Passive Packet Capture: Beyond Device Polling", http://luca.ntop.org/Ring.pdf.

Deri, L. (2000) e SUIN, S. "Effective Traffic Measurement Using ntop", IEEE Communication Magazine, v. 38, pp. 144-151.

Fraleigh, C., Diot, C., Lyles, B., Moon, S., Owerzarski, P., Papagiannaki, D. e Tobagi, F. (2001) "Design and deployment of a passive monitoring infrastructure", In Passive and active measurements workshop, Amsterdam.

Hanke J., Reitsch, A. e Wichern, D. (2001) "Business Forecasting", Prentice Hall.

Papadopouli, M., Raftopoulos, E. e Shen, H. "Evaluation of short-term traffic forecasting algorithms in wireless networks", IEEE Conference on Next Generation Internet Design and Engineering (NGI'06)", Valencia, Spain.

RFC 3272 (2002) "Overview and Principles of Internet Traffic Engineering", IETF, http://www.ietf.org/rfc/rfc3272.txt.

Yu, J. T. (2004) "Performance Evaluation on Linux Bridge", Telecommunications System Management Conference 2004, USA.

Krishnamurthy, B., Sen, S., Zhang, Y., e Chen, Y. (2003) "Sketch-based change detection: methods, evaluation, and applications", In Proceedings of the 3rd ACM SIGCOMM Conference on internet Measurement, Miami, USA.