Mineração de Dados aplicado à Gerência de Desempenho de Redes de Computadores

Pierre da Costa Viana Júnior, Cláudio Alex Jorge da Rocha, Eloi Luiz Favero

Centro de Engenharia Elétrica – Universidade Federal do Pará (UFPA)
Belém – PA – Brasil

pierre_viana@prodepa.gov.br, alex@cci.unama.br, eloi@secom.ufpa.br

Abstract. *The management of computer networks is such a hard task when it concerns the administrator, because even with the conventional tools for managing, they face decisions based on data, turning the decision making process, at times, into something doubtful and uncertain. In this context, this paper proposes an application related to the performance administration of computer networks, that uses the KDD process (Knowledge Discovery in Databases), to predict the data traffic intensity with the aim of helping the network administrator make decisions.*

Resumo. *O Gerenciamento de redes de computadores é uma tarefa árdua do ponto de vista do administrador, pois mesmo com as ferramentas de gerenciamento convencionais, estes se deparam com decisões baseadas em grandes quantidades de dados brutos, tornando a tomada de decisão muitas vezes incerta e duvidosa. Neste contexto, este trabalho propõe uma aplicação relacionada ao gerenciamento de desempenho de redes de computadores, que utilize o processo de KDD (Knowledge Discovery in Databases), para prever a intensidade de tráfego de dados, com intuito de auxiliar o gerente de redes, na tomada de decisão.*

1. Introdução

Uma rede pode existir sem mecanismos de gerenciamento, todavia seu uso pode encontrar dificuldades como congestionamento, segurança, roteamento. [ROCHA 1997]. O gerenciamento de redes é usado para controlar as atividades e monitorar os recursos da rede. O trabalho básico da gerência de rede é obter informação, extraída de grande quantidade de dados brutos, para um possível diagnóstico e execução de ações para resolução de problemas. Para alcançar estes objetivos, as funções do gerenciamento devem estar contidas em diversos componentes da rede, permitindo o diagnóstico, a prevenção e a reação aos problemas [WESTPHALL 1991, WESTPHALL 1996].

No trabalho de gerenciar uma rede de computadores existe incerteza e o uso da inteligência artificial pode ser justificado pelas vantagens que se adicionam a um sistema de gerência de redes, tais como:

- A tarefa do administrador é facilitada, promovendo um melhor desempenho, pois o sistema inteligente tende a fazer ajustes mais fino e com maior abrangência, alcançando todos os segmentos da rede;

- Com os parâmetros ajustados, o sistema se torna mais ágil, reduzindo custos e aumentando a produtividade na execução dos serviços monitorados;
- O tempo de tomada de decisão é reduzido, uma vez que o sistema notifica o gerente e propõe possíveis ações a serem executadas;

2. Processo de KDD (Knowledge Discovery in Databases)

O processo de extração de conhecimento de dados é constituído por um conjunto de etapas com a finalidade de a partir de uma base de dados em estado bruto, obter conhecimento a respeito de um determinado domínio [FAYYAD, UTHURUSAMY 2002]. A interligação deste conjunto de etapas pode ser visualizada através da figura-1 e são detalhadas a seguir.

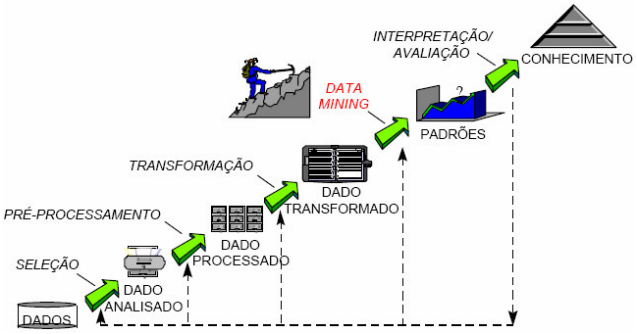

Figura 1. Etapas do Processo KDD [Fayyad et al. 1996].

2.1. Seleção

Esta etapa tem por objetivo selecionar um conjunto de dados, pertencentes a um domínio, para que, a partir de um critério definido pelo especialista, estes possam ser analisados.

2.2. Pré-processamento

Nesta etapa deverão ser realizadas tarefas que eliminem ou tratem os ruídos ou registros com dados ausentes. Outra tarefa importante é a verificação de predominância de classes, sendo que nestes casos, devem-se eliminar alguns dos registros da classe predominante ou acrescentar registros das outras classes. O objetivo é balancear a base de dados de tal forma que, no processo do aprendizado, uma classe não seja favorecida.

2.3. Transformação

Nesta etapa os dados são armazenados e formatados adequadamente para serem submetidos aos algoritmos de aprendizados.

2.4. Data Mining

Esta etapa envolve criação de modelos apropriados de representação dos padrões e relações identificadas a partir dos dados. O resultado desses modelos, depois de avaliados pelo analista e/ou especialista, são empregados para predizer os valores de atributos definidos pelo usuário final baseados em novos dados [KERBER ET AL 1995,

FAYYAD ET AL 1996B].

Neste artigo a técnica utilizada para processo de data mining foi redes bayesianas, pois eventos relacionados à utilização de redes de computadores possuem características estocásticas e probabilísticas, tornado esta a técnica mais indicada.

As redes bayesianas podem ser entendidas como modelos que codificam os relacionamentos probabilísticos entre as variáveis que representam um determinado domínio. Esses modelos possuem como componentes uma estrutura qualitativa, representando as dependências entre os nós, e quantitativa TPCs (tabelas de probabilidades condicionais desses nós), avaliando em termos probabilísticos, essas dependências [RUSSEL, NORVIG 1995].

2.5. Interpretação/Avaliação

Durante esta etapa, o conhecimento adquirido (por exemplo, árvores de decisão e regras de produção) será analisado. Para que esta análise seja feita corretamente, é fundamental que esta etapa seja realizada em conjunto com o(s) especialista(s) do domínio.

3. Gerenciamento de Redes de Computadores

Com os avanços das tecnologias de interconectividade e dos benefícios proporcionados pelas redes de computadores, cada vez mais computadores são interconectados nas organizações. Paralelamente, a diminuição dos custos dos equipamentos permite adquirir e agregar à rede cada vez mais equipamentos, de tipos diversos, tornando essas redes cada vez maiores e mais complexas. Redes locais conectadas a redes regionais, as quais, por sua vez, estão ligadas a backbones nacionais.

Este novo cenário originou alguns problemas administrativos. Tarefas antes, como configuração, identificação de falhas e controle de dispositivos da rede, passaram a ser complexas e consumir muito tempo e dinheiro.

De posse deste problema, uma área se tornou forte com intuito de solucionar ou amenizar este problema, que foi a Gerência de Redes de Computadores. Algumas definições foram propostas para a área de Gerência de redes de computadores, resumem-se a seguir algumas dessas definições [SAMPAIO 1997]:

- No controle e administração de forma racional dos recursos de hardware e software em um ambiente distribuído, buscando melhor desempenho e eficiência do sistema.
- No controle de uma rede e seus serviços. Tem por objetivo maximizar o controle organizacional das redes, de maneira mais eficiente e confiável, ou seja, planejar, supervisionar, monitorar e controlar qualquer atividade da rede.
- Tem por objetivo maximizar o controle organizacional das redes, de maneira mais eficiente e confiável, ou seja, planejar, supervisionar, monitorar e controlar qualquer atividade da rede.

4. Estudo de Caso

Existem diversas plataformas de gerência de redes no mercado, porém estas aplicações trabalham basicamente com a monitoração de variáveis e visualização gráfica das taxas

coletadas. Com isso fica totalmente a cargo do administrador a tomada de decisão, que muitas vezes se depara com situações indecisas, pois suas decisões são baseadas apenas em dados brutos.

O congestionamento é um fato que merece muita atenção, pois este comportamento vai evoluindo e caso o administrador não tome uma decisão corretiva, pode se levar muitas vezes à paralisação completa da rede.

4.1. Domínio do Trabalho

As variáveis monitoradas mostram o tráfego existente entre a rede da PRODEPA (Processamento de Dados do Estado do Pará), e rede da SESPA (Secretaria de Saúde Pública do Estado do Pará), pois são relativas à porta do roteador que serve de interface entre essas duas redes como mostra a figura-2 a seguir.

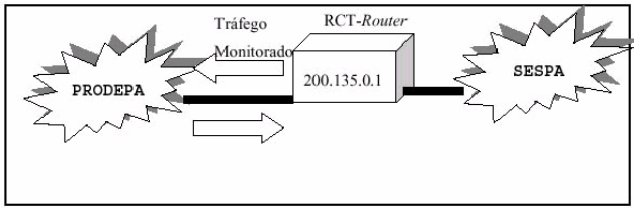

Figura 2. Tráfego monitorado

A rede da PRODEPA funciona como provedor de serviços para as mais variadas aplicações dentro do Estado. Essas aplicações variam desde sistema legados que rodam em Mainframe até a própria Internet.

O roteador monitorado é da marca Cisco, e localiza-se fisicamente nas instalações da PRODEPA. Este roteador é multiprotocolo. As interfaces de rede consistem em processadores de interfac*e* modulares, que oferecem uma conexão direta entre os barramentos de alta velocidade Cisco Extended Bus (CxBus) e uma rede externa [CISCO 1999].

4.2. As variáveis Monitoradas

As variáveis monitoradas pertencem ao grupo interfaces da MIBII - Internet especificada na RFC 1213 [IETF 1999]. O grupo Interfaces oferece dados sobre cada interface de um dispositivo gerenciável da rede. As variáveis monitoradas são:

ifOutOctets **-** O número total de bytes transmitidos por uma interface, incluindo caracteres de cabeçalho. Nome: IF-MIB!ifOutOctets; Identificador: 1.3.6.1.2.1.2.2.1.16;

ifInOctets **-** O número total de bytes recebidos em uma interface, incluindo caracteres de cabeçalho. Nome: IF-MIB!ifInOctets; Identificador: 1.3.6.1.2.1.2.2.1.10;

Estas variáveis são do tipo numérico e possuem características de contador. Suas taxas são armazenadas em bits por segundo (bps).

4.3. Metodologia de Desenvolvimento

A primeira etapa desta aplicação é monitorar um segmento da rede, coletando e gravando os dados. Paralelamente à coleta dos dados é criada uma base de dados com os valores das variáveis monitoradas. Técnicas estatísticas de mineração de dados são

aplicadas na base de dados para criar a rede bayesiana. Este trabalho culmina com a implementação de um modelo de rede bayesiana que é capaz de predizer o comportamento do segmento monitorado da rede, segundo os parâmetros selecionados na consulta. Abaixo na figura-3, é mostrado a estrutura da metodologia adotada.

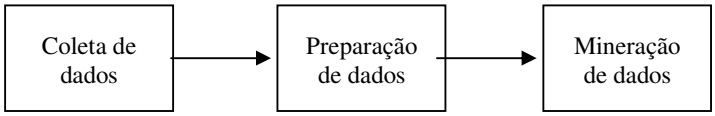

Figura 3. Metodologia de desenvolvimento.

4.3.1. Coleta de dados

Nesta etapa, é monitorado o segmento de rede de computadores proposto a cada cinco minutos e são coletados os dados do problema, além da estruturação para que possa ser submetida à etapa seguinte.

As variáveis coletadas são ifInOctets(TaxaEntrada), ifOutOctets(TaxaSaida), que são as variáveis responsáveis pelo medição do tráfego de entrada e saída respectivamente do roteador monitorado, ano, mês, dia, diasemana e hora das respectivas coletas. Em seguida é mostrada na figura-4 a estrutura da tabela implementada. Essa tabela é única no banco de dados, tornando com isso irrelevante a construção de modelos de dados para especificação.

	Nome do campo	Tipo de dados
	Ano	Texto
	Mes	Texto
	Dia	Texto
	DiaSemana	Texto
	Hora	Texto
	TaxaEntrada	Número
	TaxaSaida	Número

Figura 4. Estrutura da tabela do banco de dados.

Para implementação do programa de coleta de dados foi utilizada a linguagem orientada a objetos, Java; o conjunto de bibliotecas para gerenciamento SNMP, Adventnet e o SGBD Access da Microsoft.

O AdventNet, que possui três versões correspondentes ao protocolo SNMP, é um conjunto de bibliotecas em Java para o desenvolvimento de applets e aplicações de gerenciamento de redes usando este protocolo. Neste trabalho, é utilizada a versão AdventNet SNMP v2.

4.3.2. Preparação de dados

Esta é a etapa onde os dados são tratados, eliminando dados ausentes e ruídos, e formatados para que possam ser submetidos à ferramenta de data mining.

Os dados armazenados na tabela do banco de dados são mostrado no exemplo da figura-5 a seguir.

Ano	Mes	Dia	DiaSemana	Hora	Minuto	TaxaEntrada	TaxaSaida
2004	Janeiro	Primeiro	Quinta-Feira	Zero	Cinco	1870000	1996000
2004	Janeiro	Primeiro	Quinta-Feira	Zero	Dez	1258000	1409000
2004	Janeiro	Primeiro	Quinta-Feira	Zero	Quinze	1320000	1421000
2004	Janeiro	Primeiro	Quinta-Feira	Zero	Vinte	1406000	1582000
2004	Janeiro	Primeiro	Quinta-Feira	Zero	Vinte Cinco	1371000	1471000

Figura 5. Exemplo de registros do banco de dados.

Alguns problemas como queda do link de comunicação ou até mesmo retardo de rede, fazem com que algumas taxas sejam retornadas com valor zero. Para tratamento destes erros e ruídos encontrados nos registros, foram calculadas as médias das taxas por hora e com isso houve uma diminuição significativa no número de registros a serem submetidos à próxima fase como mostra a figura-6 a seguir.

Mes	Dia	DiaSemana	Hora	MediaTaxaEntr	MediaTaxaSaid
Abril	Décimo	Sábado	Cinco	557750,00	492750,00
Abril	Décimo	Sábado	Dez	589833,33	379500,00
Abril	Décimo	Sábado	Dezenove	599083,33	554250,00
Abril	Décimo	Sábado	Dezesseis	422583,33	611083,33
Abril	Décimo	Sábado	Dezessete	582666,67	498333,33
Abril	Décimo	Sábado	Dezoito	435666,67	469500,00

Figura 6. Exemplo das médias dos registros por hora.

Para melhor compreensão do modelo bayesiano, uma terceira variável chamada de tráfego, foi acrescentada ao problema, que depende exclusivamente da soma dos valores das duas variáveis coletadas, e o resultado da soma pode assumir uma das três classes descritas abaixo:

- Baixo – Se a soma for menor ou igual a 1200k.
- Normal – Se a soma for maior que 1200k e menor ou igual a 2400k.
- Alto – Se a soma for maior que 24000k.

Após a preparação dos dados os registros foram armazenados como mostra a figura-7.

Mes	Dia	DiaSemana	Hora	Trafego
Abril	Décimo	Sábado	Cinco	Alto
Abril	Décimo	Sábado	Dez	Normal
Abril	Décimo	Sábado	Dezenove	Alto
Abril	Décimo	Sábado	Dezesseis	Alto

Figura 7. Exemplo dos registros após a fase de preparação dos dados.

4.3.3. Mineração de dados

Nesta etapa é utilizada a ferramenta Bayesware Discoverer. Esta ferramenta utiliza técnicas para a obtenção do modelo bayesiano a partir do conhecimento implícito nos dados[BAYESWARE 2004]. Após o cumprimento desta etapa, o resultado gerado é um modelo bayesiano que contém conhecimento referente ao tráfego do segmento de rede monitorado, que possa ser utilizado no auxílio à tomada de decisão pelo administrador de redes de computadores.

Na figura-8 a seguir é mostrado o modelo bayesiano a priori, gerado a partir dos dados coletados de 01/01/2004 a 31/12/2004.

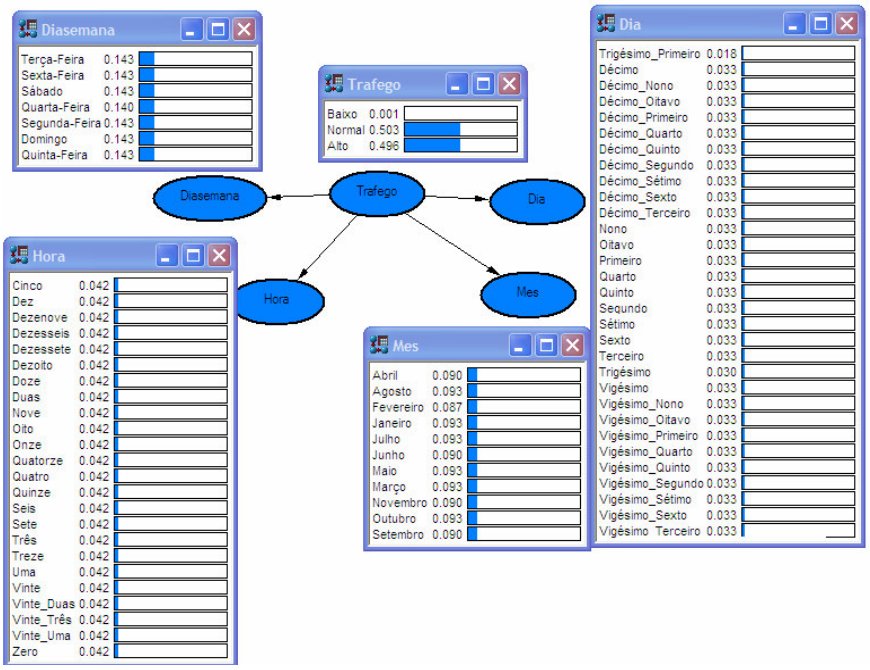

Figura 8. Modelo bayesiano a priori.

De posse deste modelo gerado, podemos fazer qualquer tipo de inferência, como mostra o exemplo a seguir, se for constatados que o dia é o Vigésimo e o Mês é Abril, como mostra figura-9 abaixo.

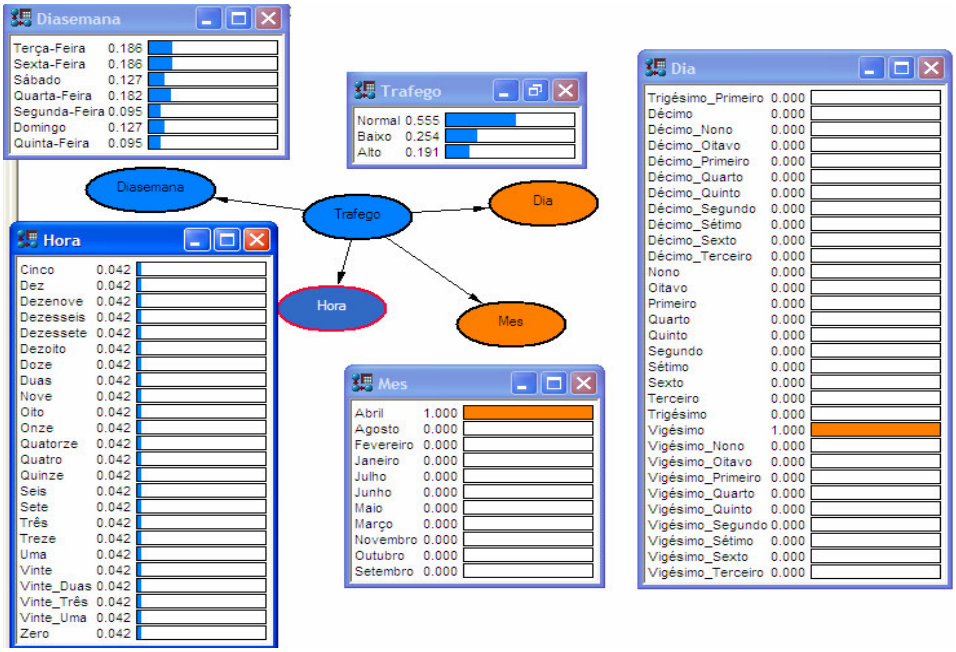

Figura 9. Modelo bayesiano após a inferência.

Temos com resposta:

A probabilidade de o tráfego ser ***intenso*** dado que o dia é o Vigésimo e o Mês é Abril é dada por: **19,1%.**

A probabilidade de o tráfego ser ***normal*** dado que o dia é o Vigésimo e o Mês é Abril é dada por: **55,5%.**

A probabilidade de o tráfego ser ***baixo*** dado que o dia é o Vigésimo e o Mês é Abril é dada por: **25,4%.**

De posse da constatação acima verificamos que o dia 20/04 é um dia propício a executarmos rotinas que requerem um fluxo de rede de média intensidade, pois como foi mostrado existe uma probabilidade de 55,5% de o fluxo ser normal nesta data.

5. Conclusão

Com os experimentos realizados e através do protótipo implementado constatou-se a adequação do enfoque probabilístico no desenvolvimento de um sistema inteligente de apoio à gerência de redes. O protótipo implementado ajuda também a compreender melhor o raciocínio sob incerteza, podendo ser usado para o treinamento de futuros administradores.

Em acréscimo, o artigo apresentou um novo conceito na área de Gerência de Redes, o conceito de "Baselines" Dinâmicas". Onde a rede bayesiana implementada é utilizada para expressar o comportamento da rede, atualizando-se com as mudanças na mesma. Uma das vantagens da utilização da baseline implementada é que ela reflete o comportamento da rede através de probabilidades, ou seja, um determinado comportamento pode ser estimado a probabilidade de sua ocorrência e verificar se estão dentro do esperado e não, como ocorre nas baselines convencionais, simplesmente estar ou não de acordo com o perfil da rede monitorada.

Referências

[BAYESWARE 2004] BWD. URL: http://www.bayesware.com, janeiro de 2005.

[CISCO 1999] CISCO. URL: http://cisco.com/warp/public/733/7000/, janeiro de 2005.

[FAYYAD ET AL 1996B] Fayyad, U.; Haussler, D.; Stolorz, P. KDD for Science Data Analysis: Issues and Examples. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), ed. Evangelos Simoudis and Jia Wei Han en Usama Fayyad, AAAI Press, pp.55-56, 1996.

[FAYYAD, UTHURUSAMY 2002] Fayyad, U. M. and Uthurusamy, R. "Evolving Data Mining into Solutions for Insights", Comminications of the ACM, vol.45, Nº 8, p. 28-31, 2002.

[IETF 1999] IETF – Internet Engening Task Force URL:http://www.ietf.cnri.reeston.va.us/rfc/rfc1213.txt, janeiro de 2005.

[KERBER ET AL 1995] Kerber, R.; Livezey, B.; Simound, E. A Hybrid System for Data Mining (Chapter 7). Itelligent Hybrid System, John Wiley & Sons Ltd, pp.121-141, 1995.

[RUSSEL, NORVIG 1995] Stuart Russel and Peter Norvig. *Artificial Intelligence, a Modern Approach*. Prentice-Hall, 1995.

[ROCHA 1997] ROCHA, M.A.; WESTPHALL, C.B. "Proactive Management of Computer Networks using Artificial Intelligence Agents and Techniques". Proceedings of the Symposium on Integrated Network Management. San Diego (CA), USA. May, 1997.

[SAMPAIO 1997] SAMPAIO, S.C. "Plataforma para concepção de aplicações de gerência utilizando o SNMP". Projeto Específico. UNIFACS– Universidade Salvador S/C. Salvador – BA, 1997.

[WESTPHALL 1991] WESTPHALL, C.B. "Conception et développement de l'architecture d'administration d'un réseau métropolitain". Thèse de doctorat nouveau régime. L' université Paul Sabatier. Toulouse, le 16 juillet 1991.

[WESTPHALL 1996] WESTPHALL, C.B.; KORMANN, L.F. "Usage of the TMN Comcepts for Configuration Management of ATM Network". International Symposium on Advanced Imaging and NetWork Technologies. Berlim, Alemanha Out. 7-11, 1996.